

Análisis de mercado inmobiliario en área metropolitana de Barcelona mediante ciencia de datos



Universitat Oberta
de Catalunya

Abel Mora Vázquez

Ciencia de datos

Máster en Ciencia de datos

Director de TFM:
Rafael Luque Ocaña

Profesora responsable de la
asignatura:

Susana Acedo Nadal

09 de enero de 2026



Esta obra está sujeta a una licencia de Creative Commons
Reconocimiento-NoComercial-SinObraDerivada 4.0
Internacional.

Para ver una copia de esta licencia, visite
<https://creativecommons.org/licenses/by-nc-nd/4.0/deed.es>

Ficha del Trabajo Final

Título del trabajo:	Análisis de mercado inmobiliario en área metropolitana mediante la ciencia de datos.
Nombre del autor/a:	Abel Mora Vázquez
Nombre del Tutor/a de TF:	Rafael Luque Ocaña
Nombre del/de la PRA:	Susana Acedo Nadal
Fecha de entrega:	01/2026
Titulación o programa:	<i>Máster de Ciencia de Datos Aplicada</i>
Área del Trabajo Final:	Ciencia de datos
Idioma del trabajo:	Castellano
Palabras clave	Machine Learning, Segmentación, Predicción
Resumen del Trabajo	
El trabajo final pretende analizar el mercado inmobiliario en el área metropolitana de Barcelona. Para ello se va a trabajar con datos históricos de precios de inmuebles en la zona, datos de características de viviendas publicadas en portales inmobiliarios, datos sociodemográficos que permitan realizar una clasificación (clustering) de los distritos para poder determinar la influencia del barrio en el precio. Una vez realizado el análisis se pretende realizar una predicción de precio en función de unas características del inmueble. Se valorará acceso a datos en vivo de plazas de parking o servicios de transporte público. También se pretende realizar una recomendación de distritos en función de ingresos o zona cercana al trabajo, así como una visualización de los datos relevantes de una vivienda incluyendo la ruta en vivo a un punto de interés del usuario ya sea punto de interés de la ciudad o centro de trabajo, para poder comparar zonas en cuanto al tiempo que requiere el desplazamiento desde la vivienda al punto seleccionado.	

Abstract

The final project aims to analyze the real estate market in the Barcelona metropolitan area. To achieve this, historical data on property prices, housing characteristics from real estate listings, and sociodemographic data will be used to perform a district classification (clustering) in order to determine the influence of location on property prices. Once the analysis is complete, the project will focus on predicting property prices based on specific housing features. Additionally, the integration of live data sources—such as parking availability and public transportation services—will be considered. The project also intends to develop a district recommendation system based on users' income or proximity to their workplace, as well as an interactive visualization of relevant housing data. This visualization will include real-time route mapping to points of interest or workplaces, allowing users to compare different areas based on travel time from a given property to the selected destination.

Índex

1.	Introducción	1
1.1.	Contexto y justificación del Trabajo	1
1.2.	Objetivos del Trabajo	2
1.3.	Impacto en sostenibilidad, ético-social y de diversidad	3
1.4.	Enfoque y método seguido	5
1.5.	Planificación del trabajo	5
1.6.	Breve sumario de productos obtenidos	9
1.7.	Breve descripción de otros capítulos de la memoria	10
2.	Materiales y métodos	11
3.	Resultados	22
4.	Conclusiones y trabajos futuros	59
5.	Glosario	62
6.	Bibliografía	63
7.	Anexos	68

Listado de Figuras

Figura 1 Relación (CCEG) con ODS	4
Figura 2 Tareas y Diagrama de Gantt	7
Figura 3 Tareas Diagrama de Gantt	8
Figura 4: Diagrama de Gantt	8
Figura 5 Evolución de precios por provincias	12
Figura 6 Guía para vivir en Barcelona	13
Figura 7 Guía ODI	19
Figura 8 Delitos contra las personas por distrito	25
Figura 9 Delitos contra el patrimonio por distrito	25
Figura 10 Otros delitos por distrito	26
Figura 11 Distribución de recursos de trenes por categoría	28
Figura 12 Radar tasas de transporte público y privado por distrito	29
Figura 13 Radar estaciones de Bicing y paradas de bus por distrito	30
Figura 14 Radar aparcamientos por distrito	19
Figura 15 Mapa de calor matriz de correlación de variables df_distritos	34
Figura 16 Mapa de calor carga variables en los componentes principales	38
Figura 17 Gráficas resultados método Elbow y método Silhouette	39
Figura 18 Grafico dispersión clústeres obtenidos Kmeans	40
Figura 19 Distribución consultas API Idealista	42
Figura 20 Distribución de habitaciones en alquiler por distrito	46
Figura 21 Precio medio de alquiler de habitaciones por distrito	47
Figura 22 Distribución de viviendas en alquiler por distrito	48
Figura 23 Precio medio de viviendas en alquiler por distrito	49
Figura 24 Distribución de precio de viviendas en alquiler por distrito	49
Figura 25 Distribución de precio de viviendas en alquiler por distrito	50
Figura 26 Distribución de viviendas en compraventa por distrito	52
Figura 27 Precio medio de viviendas en compraventa por distrito	53
Figura 28 Distribución de precio de viviendas en compraventa por distrito	53
Figura 29 Distribución precio y viviendas en compraventa	54
Figura 30 Comparación error relativo sobre precio medio de los modelos	56
Figura 31 Relación error RMSE con el tiempo de ejecución	57
Figura 32 Error de modelo por tramos de precio	58
Figura 33 Gráfica de variables más influyentes modelo CatBoost	58
Figura 34 Valor percibido frente valor precio vivienda	59
Figura 35 Arquitectura de datos y modelado (Anexo)	67
Figura 36 Mockup de visualización (Anexo)	68

Lista de Tablas

Tabla 1 Tareas	5
Tabla 2 Subtareas	6
Tabla 3 Hitos	6
Tabla 4 Gestión de riesgos	9
Tabla 5 Dataset df_poblacion	23
Tabla 6 Dataset df_superficie	24
Tabla 7 Dataset df_criminalidad	24
Tabla 8 Dataset df_movilidad	27
Tabla 9 Dataset df_bus	27
Tabla 10 Dataset df_trenes	28
Tabla 14 Dataset_df_distritos	33
Tabla 15 Descripción componentes principales	37
Tabla 16 Relación variables con los clústers	41
Tabla 17 Dataset barcelona_habitaciones	45
Tabla 18 Dataset barcelona_alquiler	47
Tabla 19 Resultado modelos de predicción precio vivienda	56

1. Introducción

1.1. Contexto y justificación del Trabajo

El trabajo final consiste en partiendo de varios conjuntos de datos, desarrollar en diferentes etapas: análisis exploratorio, limpieza de datos y adecuación a la normativa de privacidad de datos, analizar los históricos de precios por metro cuadrado de vivienda, aplicar técnicas de machine Learning para predecir los precios, realizar clustering de barrios cruzando datos sociodemográficos y datos en streaming. Finalmente crear una visualización que aporte insights relevantes para los posibles compradores o inquilinos. La justificación es el elevado precio de las viviendas, lo que dificulta el acceso a la vivienda en una zona específica. Se ha escogido el área metropolitana de Barcelona debido a que se dispone de mucha información sociodemográfica, así como ser un punto de referencia tanto estatal como internacional en diferentes sectores.

Inicialmente se pretendía realizar un proyecto de análisis de cliente 360, que pudiera aportar insights relevantes para un caso específico. Tras realizar consultas a empresas sobre la viabilidad de realizar una colaboración facilitando datos para el proyecto se ha descartado la posibilidad ya que la mayoría de las empresas contactadas han decidido no facilitar datos para el trabajo.

Actualmente estoy realizando búsqueda de vivienda en alquiler y al cambiar de zona de residencia me he percatado que no es sencillo conocer si una zona es considerada buena o mala ya que además de la opinión personal, existen variables como distancia al trabajo, tipo de población, ruidos, seguridad servicios. Por ello, se ha decidido realizar el trabajo final en esta área ya que pretende ser una ayuda para que personas no residentes en la zona tengan más información al comparar dos viviendas de precio similar.

1.2. Objetivos del Trabajo

Para determinar los objetivos se responden a las preguntas propuestas:

- ¿Qué quiero hacer?

Poner en práctica los conocimientos adquiridos durante el máster que me permita obtener una visión del sector inmobiliario en el área metropolitana de Barcelona.

- ¿Qué deseo comprobar/aportar/modificar con el TFM?

Quiero tratar de obtener información sobre barrios, población, tratar de medir la calidad de vida, realizar una predicción de precios de vivienda, un sistema recomendador de barrios en función de las características del usuario y una visualización que aporte información adicional.

- ¿Cuál es el alcance de mi trabajo (qué entra y que no)?

El alcance es tratar de dar una visión de las zonas del área metropolitana de Barcelona a un usuario interesado en alquilar o comprar una vivienda en la que pueda además de comparar precios en función de las características de la vivienda, comparar los barrios en función de los servicios, población y distancias a puntos de interés o centros de trabajo.

Los objetivos principales del trabajo son:

Predicción de precio de una vivienda dada su localización y características mediante modelos de machine Learning

- Clustering de barrios en función de la calidad de vida.
- Sistema recomendador de barrios en función de necesidades del usuario
- Visualización de servicios de movilidad y distancias a puntos de interés

Los objetivos secundarios del trabajo son:

Realizar una aplicación que permita a un usuario introducir datos de ingresos y zona de trabajo y obtener barrios recomendados. Las preguntas a las que se obtendrán respuesta son:

- ¿En qué barrio puedo residir en función de mis ingresos?
- ¿Qué barrio se adapta mejor a mis necesidades de movilidad?
- ¿Qué barrio se adapta mejor a mis preferencias?

1.3. Impacto en sostenibilidad, ético-social y de diversidad

Para analizar el impacto del trabajo en estos campos se tiene como referencia la competencia de compromiso ético y global (CCEG) que se alinean con los Objetivos de Desarrollo Sostenible (ODS) en inglés (SDG). La definición por parte de las Naciones Unidas es: *"Los Objetivos de Desarrollo Sostenible son un llamado a la acción por parte de todos los países (pobres, ricos y de ingresos medios) para promover la prosperidad y al mismo tiempo proteger el planeta. Reconocen que poner fin a la pobreza debe ir de la mano de estrategias que generen crecimiento económico y aborden una variedad de necesidades sociales, incluidas la educación, la salud, la protección social y las oportunidades laborales, al tiempo que se aborda el cambio climático y la protección ambiental (1)."* Son por tanto unos propósitos que favorecen a la sociedad y al planeta. Se analiza el impacto en las 3 dimensiones sostenibilidad, ética y diversidad.

1.3.1 Dimensión sostenibilidad

El trabajo final de máster tiene un impacto negativo medioambiental y en huella ecológica. Para el desarrollo del trabajo se ha de utilizar unos dispositivos y acceder a servidores que consumen energía, así como una vez entregado se hará uso de dispositivos de almacenamiento sea físico o en la nube que también tienen un impacto negativo. El impacto negativo no se puede cuantificar, pero sí que se ha tratado de mitigar haciendo un uso responsable de los medios utilizados para la ejecución tanto físicos como tecnologías, tratando de optimizar los procesos para que tarden el mínimo tiempo posible y con ello el menos uso de energía que puede contribuir a la contaminación. La entrega digital del trabajo contribuye a reducir su impacto medioambiental ya que no es necesaria la impresión en hojas de papel ni el uso de tinta.

1.3.2 Dimensión comportamiento ético y de responsabilidad social

El trabajo tiene un impacto en comportamiento ético y de responsabilidad social. Para mitigarlo se ha realizado siguiendo las normativas europea y española de protección de datos y siguiendo una guía de comportamiento ético tratando de no tener un impacto negativo. En cuanto a la propiedad intelectual se siguen las normativas y los usos de licencia que han establecido los autores de las referencias utilizadas. Los resultados obtenidos de

este trabajo no ponen en riesgo o empeoran algún tipo de trabajo ya que se trata de un trabajo académico no existiendo ninguna relación comercial con ninguna empresa.

1.3.3 Dimensión diversidad, género y derechos humanos

El trabajo tiene un impacto en la diversidad de género y derechos humanos negativo. Se tratan datos de personas por lo que existe la necesidad de tomar medidas que mitiguen este impacto. Las medidas tomadas son el tratamiento de los datos siguiendo las normativas europea y española de protección de datos, así como realizar un apartado que trata la perspectiva de género.

1.3.4 Objetivos de Desarrollo Sostenible (ODS)

Los ODS que se tratan en este trabajo son:

ODS 9 Industria innovación e infraestructura

ODS10 Reducción de las desigualdades

ODS 11 Ciudades y comunidades sostenibles

ODS 13 Acción por el clima

ODS 16 Paz, justicia e instituciones sólidas

En cuanto a los ODS se podrían alinear con las (CCEG) de la siguiente manera:



Figura 1 Relación (CCEG) con ODS

1.4. Enfoque y método seguido

El enfoque de este trabajo es práctico y con fines académicos ya que los datos pueden no ser replicables en otras áreas metropolitanas y no se pretende realizar un artículo científico u obtener un nuevo conocimiento que pueda ser relevante para la sociedad. El método a seguir es la realización de tareas de la planificación inicial junto con la planificación de PEC de la asignatura por lo que se trataría de una metodología Agile para poder revisar las tareas realizadas, las pendientes y adaptando la carga de trabajo en caso de ser necesario. Se realizará recogida de datos de portales con datos abiertos, análisis exploratorio de datos, analítica de barrios mediante análisis socio demográficos, clustering, reglas de asociación y predicción de precios mediante modelos de series temporales.

Para el cumplimiento normativo se tendrá en cuenta la normativa europea (LOPDGDD).

(PARLAMENTO EUROPEO, 2016.)

Para el análisis de barrios se puede realizar clustering con k-means.

Para la predicción de precio modelos de árboles como Random Forest, XGBoost...

Para la predicción de precios y series temporales algoritmo Prophet o Arima.

Para la visualización aplicaciones como Streamlit, Gradio o Tableau

1.5. Planificación del trabajo

Para la planificación se ha seguido una estructura de tareas principales que considero que deben de llevarse a cabo. Esta estructura es provisional y modificable cuando se definan tanto las tareas finales a realizar como los objetivos de las entregas. (Pérez Gómez, 2020)

Para poder realizar el calendario se considera un conjunto de tareas que van en línea con los módulos de la asignatura. Para cada tarea existen subtareas e hitos.

1.5.1 Tareas

- | |
|---|
| • Definición y planificación del trabajo final |
| • Estado del arte o análisis del mercado del proyecto |
| • Diseño e implementación del trabajo final |
| • Redacción de la documentación del trabajo final |
| • Defensa del proyecto |

Tabla 1 Tareas

1.5.2 Subtareas

- | |
|---|
| • Investigación |
| • Objetivos principales y secundarios |
| • Definir metodología |
| • Planificación del trabajo final |
| • Obtención de datos |
| • Análisis Exploratorio y limpieza de datos |
| • Análisis de la ética y legalidad de los datos, cumplimiento normativo |
| • Analítica de barrios y clustering |
| • Estudio de demanda y datos históricos |
| • Predicción de precios |
| • Sistema de recomendación |
| • Desarrollo de gráficas de visualización de los datos |
| • Creación de panel interactivo amigable |
| • Desarrollo de memoria final y presentación |
| • Desarrollo de la presentación |

Tabla 2 Subtareas

1.5.3 Hitos

- | |
|--|
| • Entrega PEC 1 |
| • Validación del conjunto de datos del trabajo final |
| • Entrega PEC 2 |
| • Finalizar técnicas de Machine Learning |
| • Finalizar visualización |
| • Entrega PEC 3 |
| • Entrega preliminar memoria |
| • Finalizar memoria |
| • Finalizar presentación |
| • Entrega PEC 4 |
| • Defensa trabajo final de máster |

Tabla 3 Hitos

1.5.4 Calendario

Para la creación del calendario se utiliza la herramienta en línea GanttPRO(2) Se ha establecido unas 20 horas semanales por tarea excepto en tareas que se estima se puedan demorar donde se han asignado dos semanas.

Para el desarrollo de la memoria final y la presentación es la tarea que más tiempo se le ha otorgado, 70 horas. En el desarrollo de la memoria se debe de incluir todos los datos obtenidos, así como el análisis del trabajo final en su totalidad y las referencias bibliográfica También la creación de presentación y la práctica de las defensas tanto en la creación del video de defensa como de la defensa ante el tribunal para lo que se dispone de los recursos de la UOC(3) y (4)

Uoc | Planificación trabajo final de máster Abel Mora Vázquez

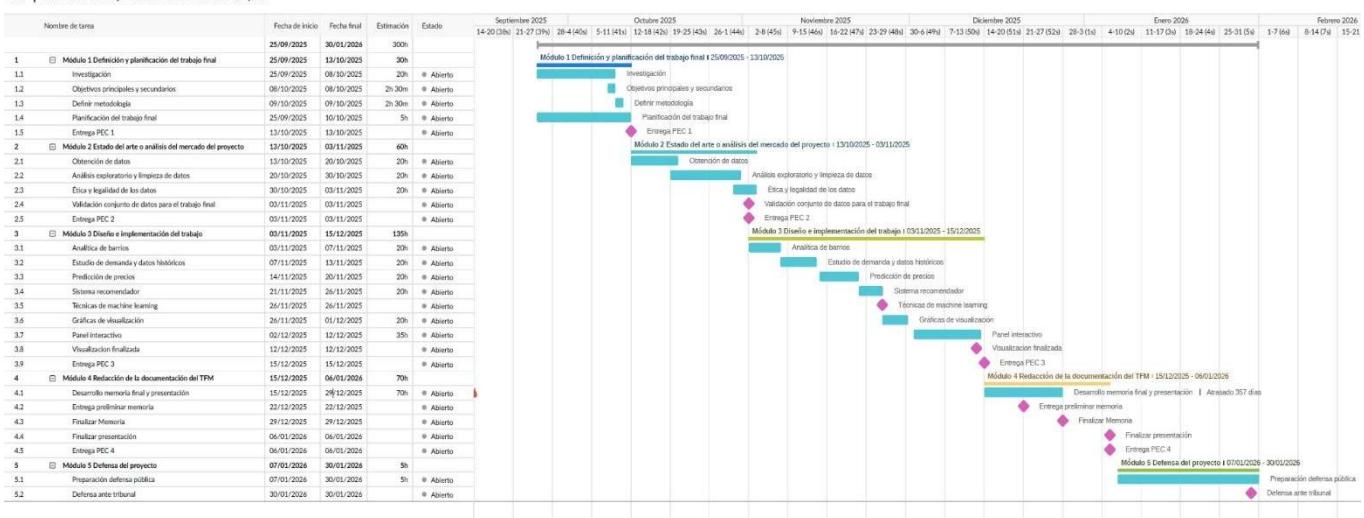


Figura 2 Tareas y Diagrama de Gantt

Nombre de tarea		Fecha de inicio	Fecha final	Estimación
1	☐ Módulo 1 Definición y planificación del trabajo final	25/09/2025	30/01/2026	300h
1.1	Investigación	25/09/2025	08/10/2025	20h
1.2	Objetivos principales y secundarios	08/10/2025	08/10/2025	2h 30m
1.3	Definir metodología	09/10/2025	09/10/2025	2h 30m
1.4	Planificación del trabajo final	25/09/2025	10/10/2025	5h
1.5	Entrega PEC 1	13/10/2025	13/10/2025	
2	☐ Módulo 2 Estado del arte o análisis del mercado del proyecto	13/10/2025	03/11/2025	60h
2.1	Obtención de datos	13/10/2025	20/10/2025	20h
2.2	Análisis exploratorio y limpieza de datos	20/10/2025	30/10/2025	20h
2.3	Ética y legalidad de los datos	30/10/2025	03/11/2025	20h
2.4	Validación conjunto de datos para el trabajo final	03/11/2025	03/11/2025	
2.5	Entrega PEC 2	03/11/2025	03/11/2025	
3	☐ Módulo 3 Diseño e implementación del trabajo	03/11/2025	15/12/2025	135h
3.1	Analítica de barrios	03/11/2025	07/11/2025	20h
3.2	Estudio de demanda y datos históricos	07/11/2025	13/11/2025	20h
3.3	Predicción de precios	14/11/2025	20/11/2025	20h
3.4	Sistema recomendador	21/11/2025	26/11/2025	20h
3.5	Técnicas de machine learning	26/11/2025	26/11/2025	
3.6	Gráficas de visualización	26/11/2025	01/12/2025	20h
3.7	Panel interactivo	02/12/2025	12/12/2025	35h
3.8	Visualización finalizada	12/12/2025	12/12/2025	
3.9	Entrega PEC 3	15/12/2025	15/12/2025	
4	☐ Módulo 4 Redacción de la documentación del TFM	15/12/2025	06/01/2026	70h
4.1	Desarrollo memoria final y presentación	15/12/2025	29/12/2025	70h
4.2	Entrega preliminar memoria	22/12/2025	22/12/2025	
4.3	Finalizar Memoria	29/12/2025	29/12/2025	
4.4	Finalizar presentación	06/01/2026	06/01/2026	
4.5	Entrega PEC 4	06/01/2026	06/01/2026	
5	☐ Módulo 5 Defensa del proyecto	07/01/2026	30/01/2026	5h
5.1	Preparación defensa pública	07/01/2026	30/01/2026	5h
5.2	Defensa ante tribunal	30/01/2026	30/01/2026	

Figura 3 Tareas Diagrama de Gantt

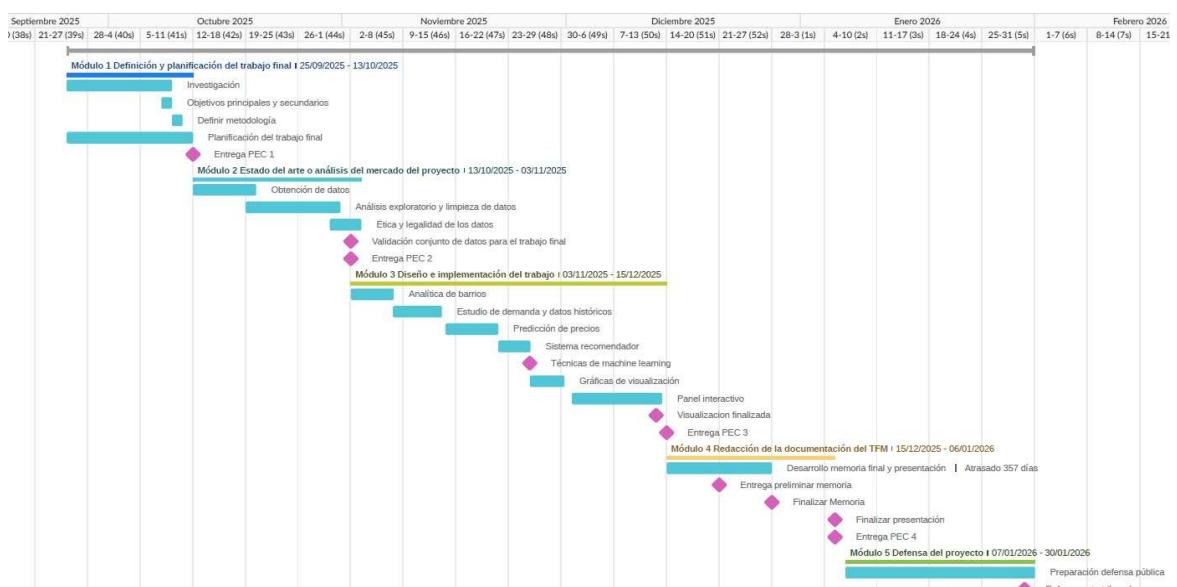


Figura 4 Diagrama de Gantt

1.5.5 Análisis de riesgos

Se han analizado los riesgos que podrían existir durante el trabajo final de máster.

Se han categorizado por tipos de riesgo internos, externos y técnicos

Se ha creado una tabla de análisis y gestión de riesgos con los riesgos analizados y las medidas a tomar para subsanarlo. La tabla es la siguiente:

Tabla de Gestión de Riesgos					
RIESGOS INTERNOS	MEDIDAS A TOMAR	RIESGOS EXTERNOS	MEDIDAS A TOMAR	RIESGOS TÉCNICOS	MEDIDAS A TOMAR
DATOS INSUFICIENTES	Consensuar con el tutor que medidas se pueden tomar para enfocar el trabajo en otra línea de investigación	MODIFICACIÓN SITUACIÓN LABORAL	Actualización de la planificación de las tareas.	FUNCIONAMIENTO INCORRECTO O ROTURA DE EQUIPOS	Copias de seguridad en discos externos o en la nube. Reparación o sustitución de los equipos o solicitud de préstamo de equipo
CUMPLIMIENTO NORMATIVO	Tratamiento de los datos con técnicas de privacidad de datos o solicitud de consentimiento	MODIFICACIÓN SITUACIÓN SALUD PERSONAL O FAMILIAR	Actualización de la planificación de las tareas y/o solicitud de aplazamiento.	TECNOLOGÍAS QUE NO FUNCIONAN	Actualización de software o sustitución del software. Aprendizaje de uso del nuevo software/técnica.

Tabla 4 Gestión de riesgos

1.6. Breve sumario de productos obtenidos

1.6.1 Plan de trabajo

El plan de trabajo se entrega en un documento que recoje toda la planificación seguida, así como las anotaciones de las correcciones que se han realizado debido a la incorrecta planificación o debido a la aparición de algunos de los riesgos planteados en el análisis de riesgos.

1.6.2 Memoria

Se entrega una memoria que recoge todos los pasos realizados en el proyecto, así como la motivación, el objetivo, los procesos realizados, los resultados obtenidos, los posibles trabajos futuros, la planificación llevada a cabo, la normativa legal tenida en cuenta, las referencias bibliográficas y los anexos.

1.6.3 Documento Notebook ipynb

Se entrega un documento ipynb que contiene el código en Python así como las anotaciones correspondientes. En el documento se incluye todos los procesos de análisis de datos, modelo de clustering de barrios, modelo de predicción de precios. También se incluirán gráficas que puedan resultar de relevancia para el trabajo.

1.6.4 Visualización

Gráficos y tablas anexas

Se han generado gráficos de los datos, así como de las planificaciones seguidas y tablas que se anexarán al documento de trabajo final de máster.

Dashboard interactivo

Se ha creado un dashboard interactivo en el que se pueden consultar los datos más relevantes del trabajo final.

1.7. Breve descripción de otros capítulos de la memoria

El capítulo 2 se describe los aspectos más relevantes del diseño y del desarrollo del trabajo.

El capítulo 3 describe los resultados que se han obtenido

El capítulo 4 recoge las conclusiones obtenidas del trabajo y los posibles trabajos futuros.

El capítulo 5 contiene el glosario de los términos y los acrónimos más relevantes utilizados.

El capítulo 6 se dedica a la Bibliografía que incluye la citación de los autores y obras de todos los recursos que se han utilizado para llevar a cabo el trabajo.

2. Materiales y métodos

Estado del arte

El trabajo final de máster trata de analizar el mercado inmobiliario de un área metropolitana en este caso Barcelona, dado que el trabajo es de tiempo limitado, no va a poder ser escalable ya que se van a utilizar datasets concretos para evaluar los barrios que están disponibles para esta ciudad. Sin embargo, se analizan otros trabajos realizados en el sector inmobiliario que tienen alcance en este proyecto.

Actualmente, el aumento de precio en la vivienda de alquiler está haciendo que se cambie la tendencia a residir en grandes ciudades por parte de familias que no dispongan de una vivienda. A esto, se une la política que se está aplicando en los bancos para acceder a una hipoteca donde normalmente se solicitan unos requisitos disponer de un ahorro del 20 por ciento del valor de la vivienda en concepto de entrada y el Banco de España recomienda que la capacidad de endeudamiento no ha de superar el 40 por ciento de los ingresos netos (5). Ante esta situación, las personas que no tienen esa capacidad de ahorro y/o la capacidad de endeudamiento no es suficiente para hacer frente a la cuota de la hipoteca, se ven obligadas a destinar su salario al pago de alquiler de vivienda. En un área metropolitana el precio se incrementa cada vez más ya que se disponen de servicios de transporte, servicios públicos de sanidad, ocio y ofertas de empleo en empresas. Un indicador clave es el Índice de Mercados Inmobiliarios Españoles (IMIE)(6).

Años atrás, la búsqueda de vivienda se realizaba en inmobiliarias, a través de conocidos, mediante búsqueda en medios de prensa escrita o mediante cartel informativo de se alquila en la vivienda. Actualmente, la búsqueda y contacto se realiza mediante portales inmobiliarios digitales como Idealista, Yaencontre, Fotocasa, Pisos.com, Habitaclia o portales especializados como el mismo Idealista o airdna.co que mediante pago se puede acceder a más datos que permiten recomendar zonas o inmuebles para la inversión. También existen portales de vivienda de alquiler de corta estancia como Airbnb. Todos estos portales hacen uso de la ciencia de datos para analizar precios, distribuciones, sistemas de recomendación, análisis de series temporales, predicción de precios que son objetivo de

este trabajo final. Dentro de Idealista se encuentra Idealista/data que ofrece gran cantidad de datos e informes periódicos como la evolución del precio de la vivienda en alquiler (7) o informes y estudios de mercado inmobiliario (8).



Figura 5 Evolución de precios por provincias Fuente Idealista/data

El mismo portal de Idealista dispone del apartado Idealista/news donde se encuentran noticias relevantes del mercado inmobiliario. El portal inmobiliario Yaencontre fue adquirido en 2020 por Idealista y también ofrece información adicional sobre barrios, guías de ciudades (9) y resúmenes estadísticos de precio tanto de alquiler como de vivienda.



Figura 6 Guía para vivir en Barcelona Fuente yaencontre

Existen más portales con funcionalidades similares, sin embargo, no dejan de ser empresas, y no reflejan el mercado real solo una imagen del mismo. Se analiza la sede electrónica del catastro del Ministerio de Hacienda del Gobierno de España (10). En ella se puede acceder a la búsqueda de inmuebles, obteniendo las características del inmueble, ubicación y el valor catastral, si bien el valor catastral, debe de ser mediante registro en la sede electrónica del catastro. También dispone de una app creada recientemente donde se pueden realizar consultas y agregar fotografías de los edificios. El valor catastral es según el catastro: “*El valor catastral es un valor administrativo fijado objetivamente para cada bien inmueble y que resulta de la aplicación de los criterios de valoración recogidos en la Ponencia de valores del municipio correspondiente. Para determinar el valor catastral de un inmueble se consideran esencialmente los siguientes componentes:*

La localización del inmueble, las circunstancias urbanísticas que afecten al suelo y su aptitud para la producción.

El coste de ejecución material de las construcciones, los beneficios de la contrata, honorarios profesionales y tributos que gravan la construcción, el uso, la calidad y la antigüedad edificatoria, así como el carácter histórico-artístico u otras condiciones de las edificaciones.

Los gastos de producción y beneficios de la actividad empresarial de promoción, o los factores que correspondan en los supuestos de inexistencia de la citada promoción.

Las circunstancias y valores del mercado valor del suelo, valor de la construcción y gastos de producción y beneficios de la actividad empresarial de promoción.

Con carácter general, el valor catastral de los inmuebles no podrá superar el valor de mercado. A tal efecto, mediante orden ministerial se ha fijado un coeficiente de referencia al mercado del 0,5 en el momento de aprobación y entrada en vigor de la ponencia. En los bienes inmuebles con precio de venta limitado administrativamente, el valor catastral no podrá en ningún caso superar dicho precio.

Los valores catastrales se pueden actualizar anualmente mediante la aplicación de coeficientes aprobados por las correspondientes Leyes de Presupuestos Generales del Estado.

Arts. 22, 23 y 32 TRLCI

Normas 10, 11, 12, 13, 16 y 20 RD 1020/93"(11)

Por un lado, tenemos los portales inmobiliarios donde hacen estudios de mercado y se publican los anuncios de venta y alquileres de viviendas y por otro lado tenemos un portal del ministerio de Hacienda donde se podría consultar el valor catastral mediante registro en la sede electrónica. Pero ¿cuál es el precio real que se paga en una compra de vivienda? Ya que pueden existir negociaciones, subidas o bajadas de precio del mercado debido a agentes externos. Recientemente el Consejo General del Notariado ha creado un portal donde se puede acceder a estadísticos reales de la venta de vivienda (12) En el portal se informa: “*¿Qué es el Portal Estadístico del Notariado?*

El Portal Estadístico del Notariado es una plataforma desarrollada por el Consejo General del Notariado que ofrece información sobre el mercado inmobiliario en España, basándose en datos reales y actualizados. Es una herramienta gratuita que te ayuda a tomar decisiones sobre la compraventa de una vivienda con seguridad y transparencia.” También se indica “Explora el precio real de la vivienda con nuestro mapa interactivo

Consulta los precios reales de compraventa firmados ante notario en cualquier punto de España. Navega, selecciona tu zona de interés y accede a estadísticas detalladas: precio por metro cuadrado, tipo de vivienda, superficie media y mucho más. Descarga gráficos comparativos para analizar la evolución del mercado con datos fiables y actualizados extraídos de la base de datos oficial del Notariado español.”. Este recurso, nos acerca algo más al precio real de las viviendas.

En cuanto a trabajos académicos se han encontrado trabajos de análisis de mercado inmobiliario en estudios de Economía y Empresa donde se analiza el uso del Big Data en empresas con casos de éxito y como se podría aplicar a una empresa pequeña para mejorar los resultados.(13) También existen estudios del mercado inmobiliario en España (14) que tal y como indica en su Resumen presentan un software que permite la realización de un complejo análisis estadístico de los datos a partir de las características de las viviendas. Otro trabajo que está vinculado a la temática es análisis y predicción del mercado inmobiliario en la Comunidad de Madrid (15)en el que se aplican diferentes modelos de predicción de precios realizando una comparación entre ellos en función de unos estadísticos.

Hasta ahora se ha encontrado información sobre precios de vivienda, y la capacidad de análisis y predicción que se obtiene aplicando la ciencia de datos, pero ¿Cómo decidimos donde vivir? En algunos casos simplemente es por el precio que se puede pagar, pero en caso de que dos viviendas tengan el mismo coste hay que decidir cual se escoge. Para ello entran variables como distancia al centro de trabajo, servicios de movilidad, servicios públicos, distancia a puntos de interés. Estas variables no solo permiten decidir a una persona cual es el mejor sitio para vivir, sino que también actúan sobre el precio de la vivienda tanto en venta como en alquiler. Se obtienen datos abiertos de puntos de interés del área metropolitana de Barcelona y datos socio demográficos del portal de datos abiertos Open Data BCN (16) y servicios de movilidad Smou (17). Estos recursos permitirán crear un dataset con el que poder calificar los barrios en función de su calidad de vida, población, servicios y punto de interés que permita al interesado decidir entre un barrio u otro. También es muy válida la opinión de residentes, se observa en el estudio Changes in the effect of energy efficiency, centrality and architectural quality on multi-family values in Barcelona 2020- 2023 (18) del cual se obtienen datos de encuesta realizada a personas residentes en los barrios donde se pregunta por diferentes características del barrio como limpieza, ruido, delincuencia.

Metodología

Las técnicas que se van a aplicar en este trabajo de Ciencia de Datos son Análisis Exploratorio y Limpieza de datos, Análisis de la ética y legalidad de los datos, Analítica de Clientes, Técnicas de clusterización, Sistema recomendador de vivienda mediante reglas,

análisis y predicción de precios y técnicas de visualización de datos. A continuación, se explican estas técnicas:

Análisis Exploratorio de Datos (EDA)

Para trabajar con los datos obtenidos se ha de seguir un proceso de análisis de los datos que permita conocer en qué formato se encuentran, adaptarlos para el uso y visualizar los datos de una manera gráfica para poder descubrir patrones y distribuciones en los datos. A este proceso se le llama Exploratory Data Analysis (EDA). Según Shirly “*El EDA Analysis o análisis exploratorio de datos es una técnica estadística que apunta a revelar estructuras subyacentes, identificar patrones o anomalías y cualquier indicio de relaciones clave que existan en un conjunto de datos o data set. El objetivo del EDA no es confirmar hipótesis, sino que se centra en generar preguntas y sus posibles direcciones para las investigaciones futuras. Para entenderlo mejor: el EDA en el Data Science es el arte de hacer preguntas más que el de buscar respuestas específicas.*” (19) Durante el proceso de EDA se realiza la limpieza de datos que consiste en modificar, sustituir o eliminar datos para el uso posterior de los mismos. La limpieza de datos se ha de realizar tras analizar las consecuencias que implica cada técnica ya que se podría modificar la estructura general de los datos y obtener resultados erróneos.

Análisis de la ética y legalidad de los datos

Los datos han de cumplir la normativa vigente de protección de datos. El diccionario panhispánico jurídico define la protección de datos así: “*1. Adm. Conjunto de medidas para garantizar y proteger los datos de carácter personal (cualquier información concerniente a personas físicas identificadas o identificables) registrados en soporte físico, que los haga susceptibles de tratamiento, y a toda modalidad de uso posterior de estos datos por los sectores público y privado, a los efectos de garantizar y proteger las libertades públicas y los derechos fundamentales de las personas físicas, y especialmente de su honor e intimidad personal y familiar. Tales medidas se basan en los principios de calidad de los datos, el derecho de información en la recogida de datos, el consentimiento del afectado, los datos especialmente protegidos, los datos relativos a la salud, la seguridad de los datos, el deber de secreto, la limitación a la comunicación y el acceso a los datos por parte de terceros; así*

como en los derechos de las personas a la impugnación de valoraciones, a la consulta del Registro General de Protección de Datos, a la oposición, acceso, rectificación o cancelación de sus datos, a la tutela de tales derechos y a la indemnización. Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE; Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y Garantía de los Derechos Digitales, de España". (20). Los datos que se protegen son de carácter personal, la web de la unión europea define

los datos personales de la siguiente manera: "Los datos personales son cualquier información relacionada con una persona identificada o identifiable, también denominada "el interesado". Ejemplos de datos personales:

- nombre y apellidos
- dirección
- número de documento de identidad/pasaporte
- ingresos
- perfil cultural
- dirección de protocolo internet (IP)
- datos en poder de hospitales o médicos (que identifican únicamente a una persona con fines sanitarios).
- Categorías especiales de datos
- No se pueden tratar datos personales sobre:
 - origen racial o étnico
 - orientación sexual
 - opiniones políticas
 - convicciones religiosas o filosóficas
 - afiliación sindical
- datos genéticos, biométricos o sanitarios, salvo en casos específicos (por ejemplo, cuando se da un consentimiento explícito o cuando el tratamiento es necesario por razones de interés público esencial, sobre la base del Derecho nacional o de la UE)
- condenas e infracciones penales, a menos que lo autorice el Derecho nacional o de la UE" (21)

Que los datos cumplan con los requerimientos legales no implica que se estén tratando éticamente. Un ejemplo puede ser realizar una encuesta y no recoger datos de personas debido a su nacionalidad, sexo, raza o ideología. Gartner define la ética de los datos como: “*un sistema de valores y principios morales relacionados con la recopilación, el uso y el intercambio responsable de datos. Las violaciones a la ética de los datos van desde abiertas y públicas hasta sutiles y secretas, como algoritmos que sugieren tasas de interés más altas para los solicitantes de hipotecas de minorías o líneas de crédito más bajas para las mujeres solicitantes de tarjetas de crédito*”.) (22) Por tanto, no parece sencillo gestionar los datos de manera ética desde su creación o recopilación hasta cumplir el ciclo de vida de los datos, sin embargo, existen algunos recursos que aportan soporte en la gestión ética de los datos. La página del ministerio para la transformación digital (23) muestra varias formas en la que tratar los datos de manera ética y responsable y propone un recurso del Open Data Institute (24) que incluye un curso y una guía de buenas prácticas en la gestión ética de los datos.



Figura 7 Guía ODI

La normativa aplicable en cuanto a protección de datos es la normativa europea (25) y la española (6).

Analítica de Clientes

Para cualquier empresa resulta fundamental recopilar datos de los clientes ya sean datos demográficos, transacciones, datos de consumo, opiniones o valoraciones que haya realizado. Las empresas, cumpliendo la normativa vigente, pueden decidir qué datos recopilan y de qué manera. Para ello es necesario un estudio de los datos que se pueden recoger del cliente, los medios por los cuales se van a recoger, el formato y el tratamiento que se va a dar a los datos. Según los datos, las empresas pueden establecer estrategias de marketing, predicción de demanda, actualización de servicios o productos. La analítica de clientes permite a las empresas tomar decisiones informadas y tratar de diferenciarse de la competencia. Existen diferentes herramientas para recopilar información de los clientes, IBM (26) sugiere los siguientes:

Cookies

Paneles de control de CRM

Correo electrónico

Redes sociales

Encuestas

Sitios web²⁶

Una vez recopilados los datos se deben de tratar legal y éticamente previo a que se pueden aplicar técnicas de aprendizaje automático (machine learning) para obtener información sobre los comportamientos de compra de diferentes grupos de clientes. (27) El término aprendizaje automático o machine learning procede de Arthur Samuel (1901 – 1990) quien durante su trabajo en IBM dedicó su tiempo a desarrollar un programa Samuel Checkers el cual, era un juego de damas que trataba de ganar a un humano. Debido a los problemas de memoria, Samuel introdujo un sistema de puntuación que permitía calcular las probabilidades de ganar en ciertas posiciones. Cada vez que se jugaba una nueva partida, el sistema mejoraba y aprendía nuevas probabilidades de victoria (28). El machine learning puede ser ejecutado de 3 maneras diferentes (29):

- Aprendizaje supervisado: Los datos se dividen en conjunto de entrenamiento y prueba y se evalúa el rendimiento.

- Aprendizaje no supervisado: Los datos se usan al completo siendo el algoritmo quien detecta diferentes grupos o patrones que no se aprecian.
- Aprendizaje semi supervisado: se utilizan datos de entrenamiento con etiquetas y sin etiquetas

También existe el aprendizaje por refuerzo, el cual AWS lo califica como una técnica de machine learning: *“El aprendizaje por refuerzo (RL) es una técnica de machine learning (ML) que entrena al software para que tome decisiones y logre los mejores resultados. Imita el proceso de aprendizaje por ensayo y error que los humanos utilizan para lograr sus objetivos. Las acciones de software que trabajan para alcanzar su objetivo se refuerzan, mientras que las que se apartan del objetivo se ignoran”*(30). En la analítica de clientes se puede obtener información sobre el comportamiento de compra de los clientes, agruparlos por consumos similares, analizar que supone el cambio de precio en un producto para los clientes, analizar el ciclo de vida del cliente, es decir, el gasto total que se espera de un cliente en la empresa. En este trabajo se va a aplicar la analítica de clientes adaptada al usuario que está buscando una vivienda, en base a unas características del cliente se le clasificará y se obtendrán los barrios más adecuados mediante un sistema recomendador. Además del clúster de cliente, se va a realizar un clustering de barrios mediante el modelo no supervisado Kmeans. Su funcionamiento según explica scikit-learn: *“El algoritmo KMeans agrupa los datos intentando separar las muestras en n grupos de varianza igual, minimizando un criterio conocido como inercia o suma de cuadrados dentro del grupo (ver más abajo). Este algoritmo requiere que se especifique el número de grupos. Se adapta bien a grandes cantidades de muestras y se ha utilizado en una amplia gama de áreas de aplicación en muchos campos diferentes. El algoritmo k-means divide un conjunto de N muestras X en K cúmulos disjuntos C, cada uno descrito por la media μ_j de las muestras del conglomerado. Las medias se denominan comúnmente “centroídes” del conglomerado; tenga en cuenta que, en general no son puntos de X, aunque vivan en el mismo espacio. El algoritmo K-means tiene como objetivo elegir centroídes que minimicen la inercia o el criterio de suma de cuadrados dentro del grupo del grupo”*(31)

$$\sum_{i=0}^n \min_{\mu_j \in C} \left(\left\| x_i - \mu_j \right\|^2 \right)$$

Predicción de precios

La predicción de precio de la vivienda se va a predecir con diferentes modelos los cuales se compararán para determinar cual ofrece mejor predicción en base a métricas MAE, MSE, RMSE (32). Los modelos podrían ser Regresión lineal múltiple(33), Random Forest (34), XGBoost (35) o redes neuronales(36).

Lenguaje y entorno

En cuanto al lenguaje de programación para el trabajo final se va a utilizar principalmente Python (37) que es un lenguaje muy utilizado en Ciencia de Datos ya que permite mediante la carga de librerías realizar cálculos matemáticos y análisis de datos. Para el entorno se va a utilizar JupyterNotebook que es una aplicación web de código abierto la cual permite crear y compartir código en diferentes lenguajes de programación. Databricks define JupyterNotebook así (38) “*Un Jupyter Notebook es una aplicación web de código abierto que permite a los científicos de datos crear y compartir documentos que incluyen código en vivo, ecuaciones y otros recursos multimedia. ¿Para qué se utilizan los Jupyter Notebooks? Los cuadernos Jupyter se utilizan para todo tipo de tareas de ciencia de datos, como análisis exploratorio de datos (EDA), limpieza y transformación de datos, visualización de datos, modelado estadístico, aprendizaje automático y aprendizaje profundo*”.

Visualización de datos

Para la visualización de datos se emplearán diferentes tipos de gráficos con el fin de que describan los datos de manera visual se utilizarán librerías como Matplotlib (39) para gráficos estáticos y Seaborn (40) y Flourish (41) para gráficos elaborados. También se pretende crear una visualización interactiva mediante una app donde el usuario vea las características de la vivienda y barrio recomendado para ello es necesario un aprender a cómo usar la herramienta y la integración final con los datos. Aplicaciones que se tienen en cuenta con Streamlit (42) Gradio(43) y Tableau(44).

3. Resultados

Para la elaboración del trabajo se han obtenido datos de diferentes fuentes tanto de estamentos públicos como de empresas privadas. Los datos que se tratan en este trabajo cumplen con las licencias que han declarado los autores, citando la obra y los autores en la sección de Bibliografía.

El trabajo se ha dividido en 2 partes diferenciadas dadas las técnicas y datos necesarios:

1. Clasificación de distritos de Barcelona
2. Análisis de mercado y predicción de precios de viviendas.

Para la clasificación de distritos de Barcelona se obtienen datos del portal de datos abiertos Open Data BCN del Ajuntament de Barcelona. Para esta primera parte del trabajo se necesitan datos que describan los barrios, como datos de población, superficie, servicios públicos, datos de movilidad, datos de criminalidad, datos sociodemográficos como renta media o nivel de estudios. Con estos datos se pretende crear un dataset que describa los barrios para poder realizar una clasificación en función de sus características. Se crearán índices de tasas que reflejen la realidad del barrio como tasa de movilidad o tasa de criminalidad, así como otros estadísticos que permitan poder clasificar los barrios sin que factores como la extensión o la población influyan en los resultados.

Una vez que se ha profundizado en el trabajo, se ha debido de subir un nivel de granularidad de los datos y en lugar de hacer la clasificación de los 73 barrios como estaba previsto se ha debido de realizar el estudio de los 10 distritos. Como punto positivo, permite hacer un análisis más global ya que a la hora de clasificar un barrio, una calle puede delimitar clasificar una finca en el barrio que le pertenece pero que estadísticamente sea más cercana al barrio contrario. Se han recogido 6 conjunto de datos del repositorio de datos abiertos de Barcelona para poder conformar datasets que permitan realizar la clasificación de distritos. Estos datos se guardarán en un conjunto de datos nuevo llamado df_distritos. El dataset df_distritos se crea para concentrar los datos obtenidos de fuentes oficiales en este caso del Ajuntament de Barcelona a través del portal de datos abiertos Open Data. En este dataset se irán incluyendo los datos que se necesitan para el trabajo.

Se realiza el Análisis Exploratorio de datos de los conjuntos de datos de población, superficie, criminalidad, movilidad, servicio de buses y servicio de trenes.

3.1 Clasificación de distritos Barcelona

3.1.1 Análisis exploratorio

Dataset df_población

El dataset df_poblacion (45) contiene una serie temporal con los datos de población de los distintos distritos de Barcelona. El dataset contiene 84 registros y 137 columnas. No existen datos nulos ni duplicados. No se va a trabajar con todas las variables ya que solo interesa tener los datos de población más recientes de julio de 2025.

Nombre dataset	Número de filas			Número de columnas
df_poblacion	84			137
Nombre columna	Tipo de dato	Valores únicos	Valores nulos	Descripción
Territorio	object	84	0	Nombre del territorio
Tipo de territorio	object	4	0	Tipo de territorio
Resto de columnas	Int64	84	0	Número habitantes
Acciones llevadas a cabo				
Se filtra por tipo de territorio igual a distrito y se selecciona el dato más actualizado correspondiente a julio de 2025. Sant Andreu aparece como barrio y como distrito. Se filtran los datos de distrito. Se extraen los datos y se guardan en df_distritos.				

Tabla 5 Dataset df_poblacion

Dataset df_superficie

El dataset df_superficie (46) contiene la superficie en ha de los distritos de Barcelona del año 2021, dado que los municipios no suelen cambiar su superficie se acepta estos datos como válidos. Contiene 73 registros y 6 columnas. No existen valores nulos.

Nombre dataset	Número de filas			Número de columnas
df_superficie	73			6
Nombre columna	Tipo de dato	Valores únicos	Valores nulos	Descripción
Any	Int64	1	0	Año de los datos
Codi_Disricte	Int64	10	0	Codificación territorio
Nom_Disricte	object	10	0	Nombre de Distrito
Codi_Barri	Int64	73	0	Codificación del barrio
Nom_Barri	object	73	0	Nombre del barrio
Superficie	Float64	73	0	Superficie en ha
Acciones llevadas a cabo				
Se selecciona el código de distrito, nombre de distrito y se agrupa la superficie por distrito para obtener la suma de la superficie de los distritos agrupados en distritos. Se transforma la variable Superficie de ha a km ² . Se añaden los datos de código de distrito y superficie en km ² al dataset df_distritos.				

Tabla 6 Dataset df_superficie

Dataset df_criminalidad

El dataset df_criminalidad (47) contiene 255 registros y 7 columnas. Si bien tiene un subapartado que describe las columnas. Esto es debido a que el dataset original contiene una serie temporal. Se ha filtrado antes de descargar por los datos más recientes de diciembre de 2025. Por lo que los datos no estarán actualizados en enero de 2026

Nombre dataset	Número de filas			Número de columnas
df_criminalidad	255			7
Nombre columna	Tipo de dato	Valores únicos	Valores nulos	Descripción
Territorio	Object	12	0	Nombre de Territorio
Tipo de territorio	Object	2	0	Tipo de territorio
Categoría	object	26	0	Nombre del delito
2025	Object	2	0	Es debido al filtro n/a
2025.1 Hechos contra las personas	object	65	0	Nº de delitos de la tipología delitos contra las personas
2025.2 Hechos contra el patrimonio	object	79	0	Número de delitos de la tipología delitos contra el patrimonio
2025.3 Otros hechos	object	66	0	Número de otros delitos
Acciones llevadas a cabo				
Se agrupan las columnas por tipo de delito renombrando las columnas a Delito_Personas, Delito_Patrimonio y Delito_Otros. Se eliminan los registros que contienen como nombre no informat. Se elimina la columna 2025. Se agrupa por distrito y categoría. Se convierten las columnas a numéricas y se suman los registros de delitos. Se crean visualizaciones que representan los tipos de delito en cada distrito mediante un radar chart. Se añaden los datos al df_distritos				

Tabla 7 Dataset df_criminalidad

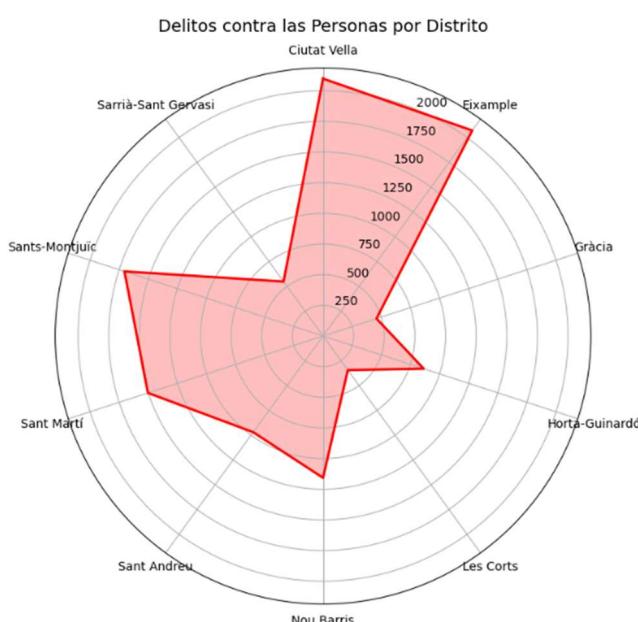


Figura 8 Delitos contra las personas por Distrito

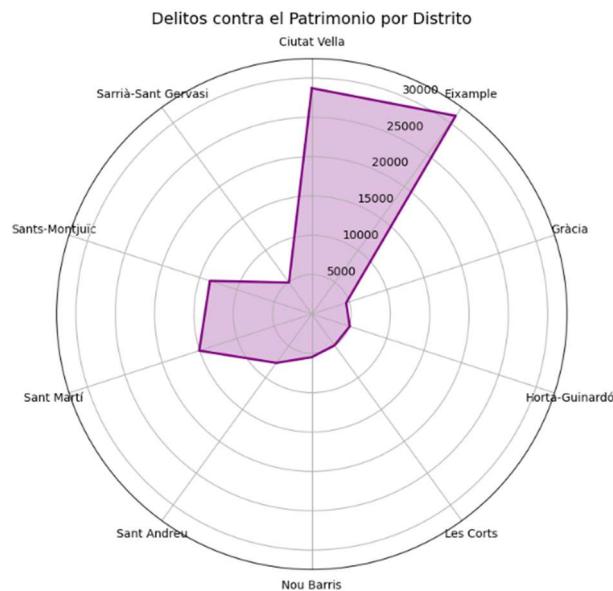


Figura 9 Delitos contra el patrimonio por Distrito

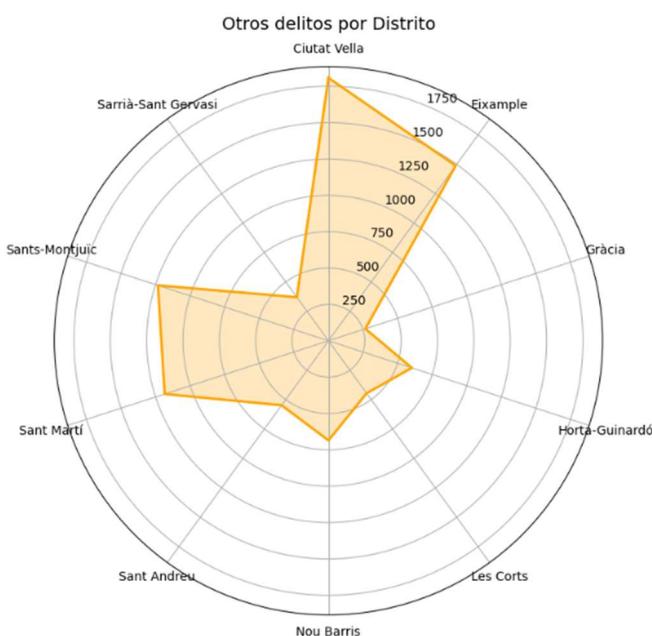


Figura 10 Otros Delitos por Distrito

En los gráficos se puede observar como en el centro del radar estaría el punto 0 y a lo largo de la circunferencia están los distritos. El número de delitos cometidos en cada distrito se proyecta en el radio que une el distrito con el punto 0. Permite ver de manera muy visual los distritos con menor y mayor número de delitos en cada categoría.

En este momento se calcula la tasa de criminalidad que ayudará a poder clasificar los distritos con un parámetro que refleje el nivel de delincuencia de cada distrito. Se ha decidido

obtener una tasa para cada tipo de delito, es decir, una tasa de criminalidad para los delitos contra las personas Tasa_C_Personas, una tasa de criminalidad para delitos contra el patrimonio Tasa_C_Patrimonio. y una tasa de criminalidad para otros delitos Tasa_C_Otros. En tipo de delito otros se incluyen delitos contra la seguridad vial. El cálculo de las tasas se ha realizado dividiendo el número de delitos de cada tipo entre el número de habitantes del distrito y multiplicado por 1000 obteniendo la tasa de criminalidad por cada 1000 habitantes.

$$\text{Tasa_C_Personas} = (\text{Delito_Personas} / \text{Poblacion}) * 1000$$

$$\text{Tasa_C_Patrimonio} = (\text{Delito_Patrimonio} / \text{Poblacion}) * 1000$$

$$\text{Tasa_C_Otros} = (\text{Delito_Otros} / \text{Poblacion}) * 1000$$

El valor de la tasa de criminalidad se añade al df_distritos.

Dataset df_movilidad

El dataset de movilidad (48) contiene muchísima información. La estructura del dataset es diferente a un dataset normal ya que incluye filtros secundarios. Dadas las características del dataset solo se mostrará en la memoria los datos que se han recogido del mismo.

Nombre dataset	Número de filas			Número de columnas
df_movilidad	7943			40
Nombre columna	Tipo dato	Valores únicos	Valores nulos	Descripción
Addresses_district_name	Object	10	0	Nombre de distrito
Secondary_filters_name	Object	4	0	Nombre de elemento de movilidad
Count	Int64	25	0	Número de elementos
Acciones llevadas a cabo				
Se han filtrado del dataset original elementos de movilidad destacables, los distritos a los que pertenecen y el número de elementos en cada barrio. Se ha creado un dataset específico para las estaciones de Bicing con la estructura de Territorio, nombre de elemento y número de elementos. Posteriormente se le ha añadido el código de distrito y se ha añadido a un dataset de movilidad df_movilidad_total.				

Tabla 8 Dataset df_movilidad

El dataset df_movilidad_total creado va a permitir unificar en un dataset datos provenientes de diferentes conjuntos de datos para posteriormente procesarlos.

Dataset df_bus

El dataset df_bus (49) contiene 3226 registros y 16 filas con datos de las paradas de autobús tanto por barrio como en diferentes codificaciones de geolocalización. Se describen en la memoria los datos que se extraen del mismo

Nombre dataset	Número de filas			Número de columnas
df_bus	7943			40
Nombre columna	Tipo de dato	Valores únicos	Valores nulos	Descripción
DISTRICTE	Object	10	0	Codificación distrito
NOM_DISTRICTE	Object	10	0	Nombre de distrito
NOM_CAPA	Object	4	0	Tipo de servicio bus
Count	Int64	23	0	Nº de servicios por distrito
Acciones llevadas a cabo				
Se ha seleccionado los tipos de servicio de bus diferentes existentes y se han extraído el distrito, el tipo de servicio y el número de elementos por cada distrito. Posteriormente se ha añadido el código y distrito y se adjuntan los datos a df_movilidad_total				

Tabla 9 Dataset df_bus

Dataset df_trenes

El dataset df_trenes (50) tiene la misma estructura que el df_bus con varias columnas de geolocalización en diferentes formatos. Se han llevado a cabo las mismas acciones. Se detallan a continuación únicamente los datos extraídos

Nombre dataset	Número de filas			Número de columnas
df_trenes	684			16
Nombre columna	Tipo de dato	Valores únicos	Valores nulos	Descripción
DISTRICTE	Object	10	0	Codificación distrito
NOM_DISTRICTE	Object	10	0	Nombre de distrito
NOM_CAPA	Object	8	0	Tipo de servicio tren
Count	Int64	18	0	Nº de servicios por distrito
Acciones llevadas a cabo				
Se ha seleccionado los tipos de servicio de trenes diferentes existentes y se han extraído el distrito, el tipo de servicio y el número de elementos por cada distrito. Posteriormente se ha añadido el código y distrito y se adjuntan los datos a df_movilidad_total				

Tabla 10 Dataset df_trenes

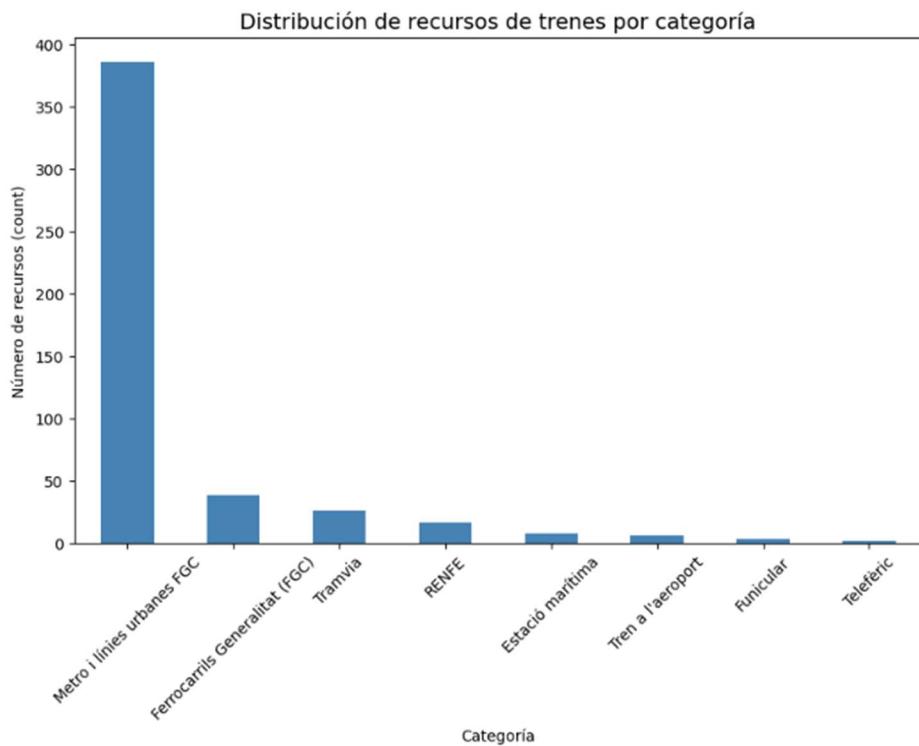


Figura 11 Distribución de recursos de trenes por categoría

Se ha decidido eliminar como elemento de movilidad el funicular y el teleférico ya que por sus características no son replicables a todos los distritos y su representación es muy pequeña. Estación marítima se mantiene ya que se refiere a la estación de tren. Se ha decidido integrar Ferrocarrils, RENFE, Estació marítima i Tren a l'aeroport dentro de una misma variable llamada tren. Por tanto, el dataset que se cargará en df_movilidad_total tendrá Metro, Tren y Tranvia como variables representativas de este dataset.

A continuación, de manera similar a la tasa de criminalidad se ha decidido crear la tasa de movilidad. En este caso se establece la tasa de movilidad por superficie Tasa_Movilidad_Sup. Para conseguir esta tasa se han agrupado todos los elementos de movilidad por distrito y se ha dividido entre la superficie del distrito. De esta manera, se obtiene la tasa de movilidad por km² del distrito no perjudicando a distritos más pequeños frente a los más grandes. También se ha creado la tasa de movilidad de transporte público y la tasa de movilidad de transporte privado. Los elementos de transporte público son tren, tranvía, metro, bus, paradas de taxi y estaciones de bicing. Los elementos de transporte privado son Puntos de recarga de automóvil eléctrico, Aparcamientos y gasolineras.

De esta manera se obtienen 3 tasas para clasificar a los distritos en cuanto a la movilidad.

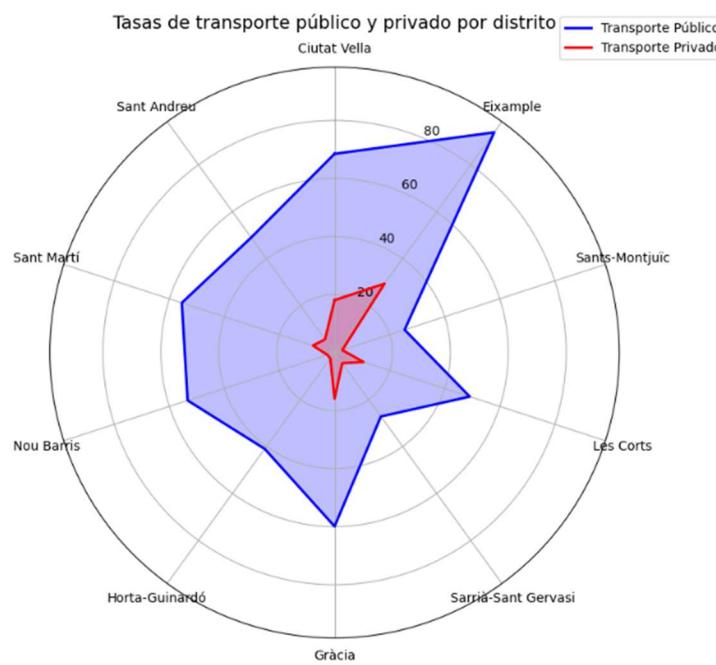
$$\text{Tasa_Movilidad_Sup} = \text{Total_Movilidad} / \text{Superficie (km}^2\text{)}$$
$$\text{Tasa_Trans_Publico} = \text{Trans_P\xfublico} / \text{Superficie (km}^2\text{)}$$
$$\text{Tasa_Trans_Privado} = \text{Trans_Privado} / \text{Superficie (km}^2\text{)}$$


Figura 12 Radar Tasas de transporte p\xfublico y privado por distrito

En el gráfico de radar podemos ver en azul la tasa de transporte público por cada distrito. Se observa como existen distritos mejor comunicados respecto a otros. También podemos observar como la tasa de transporte privado es mucho menor, esto es debido a la diferencia que existe en este momento antes de normalizar datos entre las dos variables. Por este motivo es fundamental cuando se va a realizar una clasificación de este tipo normalizar las variables para que unas no tengan más peso debido a la escala.

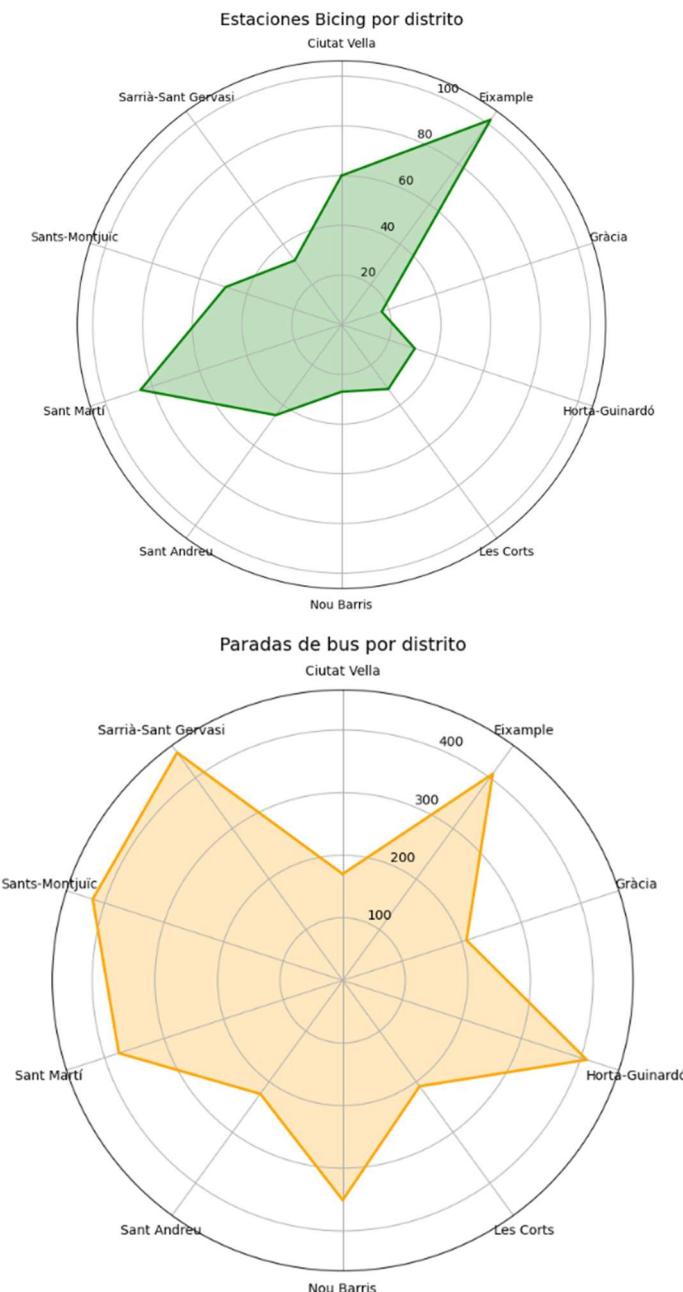


Figura 13 Radar Estaciones de Bicing por distrito y Paradas de bus por distrito

Se puede observar cómo hay distritos que tienen más recursos de un servicio de movilidad por ejemplo Nou Barris no tiene casi presencia de Bicing pero sí de paradas de bus. En cuanto a la movilidad de transporte privado, se observa cómo hay distritos que disponen de más oferta de aparcamientos que otros, algo a valorar si se dispone de un coche.

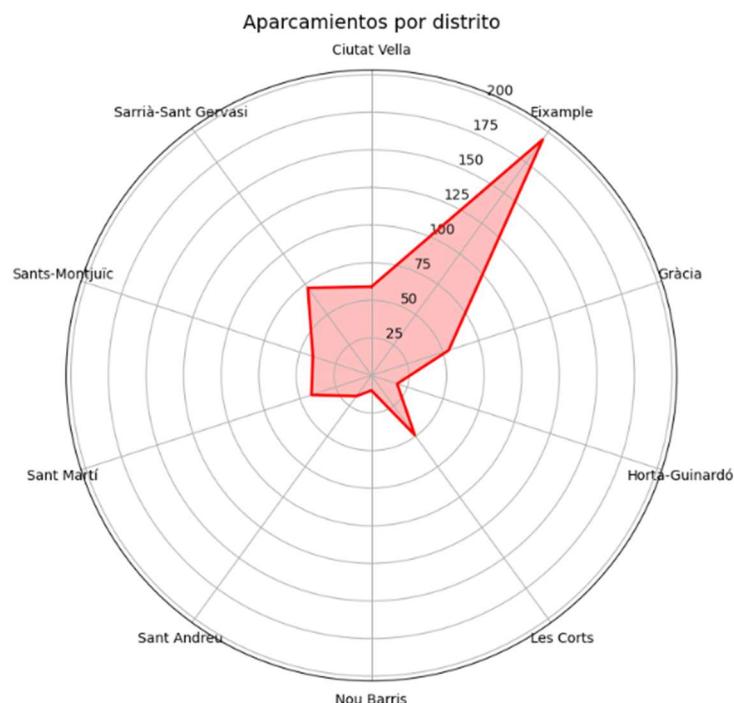


Figura 14 Radar Aparcamientos por distrito

El siguiente conjunto de datos pertenece al *Informe de Resultats de l'Enquesta de Serveis Municipals 2025* del Ajuntament de Barcelona que realiza a los vecinos (51). Es un documento que se actualiza anualmente y que contiene muchísima información tanto de población, como de valoración de servicios del ayuntamiento, intención de voto. En este caso, se ha decidido estudiar los informes que tienen publicados de la encuesta y se ha creado un nuevo dataset df_distrito. En él, se han ido añadiendo los datos que se han considerado que podían ser útiles para la clasificación de distritos. En la encuesta, los vecinos pueden valorar del 1 al 10 su grado de satisfacción con servicios del ayuntamiento, pero también pueden valorar el estado del barrio en diferentes áreas como por ejemplo en cuanto a seguridad, ruidos, limpieza. La encuesta ha tenido en cuenta 6000 valoraciones y los resultados han sido ponderados para equilibrar las muestras de los distritos mediante pesos. De estos informes se han extraído las valoraciones por distritos y se han juntado en el dataset mencionado. Las variables seleccionadas son:

Variable	Gráfico	Descripción
recodiga_basura	Recollida d'escombraries	Valoración media recogida de basuras (0-10)
limpieza_calles	Neteja dels carrers	Valoración media recogida de basuras (0-10)
ruido	Soroll	Valoración media ruido percibido (0-10)
circulacion	Circulació	Valoración media recogida de circulación (0-10)
aparcamiento	Aparcament	Valoración media facilidad de aparcamiento (0-10)
autobus	Autobus	Valoración media servicio autobús (0-10)
metro	Metro	Valoración media servicio metro (0-10)
tranvía	Tranvia	Valoración media servicio tranvía (0-10)
seguridad_barrio	Seguretat Ciutadana	Valoración media de la seguridad ciudadana (0-10)
cambio_pasado	Diferencia entre ha mejorado o ha empeorado el barrio el último año Valoración ha mejorado – Valoración ha empeorado	Índice neto del cambio percibido. Valor positivo ha mejorado valor negativo ha empeorado
cambio_futuro	Diferencia entre mejorará o empeorará Valoración mejorará – Valoración empeorará	Índice neto de expectativa futura del barrio. Valor positivo mejorará valor negativo empeorará
prob_inseguridad	Problema mes greu del barri (%) personas	% personas que indican el problema principal del barrio
prob_limpieza	Problema mes greu del barri (%) personas	% personas que indican el problema principal del barrio
prob_acceso_vivienda	Problema mes greu del barri (%) personas	% personas que indican el problema principal del barrio
prob_ruido	Problema mes greu del barri (%) personas	% personas que indican el problema principal del barrio
prob_turismo	Problema mes greu del barri (%) personas	% personas que indican el problema principal del barrio
prob_civismo	Problema mes greu del barri (%) personas	% personas que indican el problema principal del barrio
prob_transporte	Problema mes greu del barri (%) personas	% personas que indican el problema principal del barrio
prob_mantenimiento	Problema mes greu del barri (%) personas	% personas que indican el problema principal del barrio
prob_aparcamiento	Problema mes greu del barri (%) personas	% personas que indican el problema principal del barrio
prob_comercio	Problema mes greu del barri (%) personas	% personas que indican el problema principal del barrio
prob_inmigracion	Problema mes greu del barri (%) personas	% personas que indican el problema principal del barrio
prob_urbanismo	Problema mes greu del barri (%) personas	% personas que indican el problema principal del barrio

barrio_mejor_que_bcn	El barri es un dels millors respecte a Barcelona	% personas que consideran su barrio entre los mejores
barrio_peor_que_bcn	El barri es un dels pitjors respecte a Barcelona	% personas que consideran su barrio entre los peores
trabajadores_total	Situació laboral	Índice de población ocupada cuanta propia + ajena
educ_basico	Estudis	% población estudios obligatorios finalizados + secundaria general
educ_fp	Estudis	% población estudios grado medio + grado superior
educ_univ	Estudis	% población estudios universitarios
nacionalidad_esp	Nacionalitat	Suma de nacionalidad española de siempre + adquirida
nacionalidad_ue	Nacionalitat	Población nacionalidad de países UE
nacionalidad_no_ue	Nacionalitat	Población con nacionalidad fuera UE
propiedad	Regim tinença de l'habitatge	% de hogares en propiedad
alquiler	Regim tinença de l'habitatge	% de hogares en alquiler
coche_si	Bens de consum	% hogares con coche
moto_si	Bens de consum	% hogares con moto
ingresos_mensuales	Ingresos mensuales	Ingresos medios mensuales
precio_m2_alquiler	Fuente Idealista	Coste medio mensual m ² de alquiler
precio_m2_compra	Fuente Idealista	Precio medio m ² de compraventa

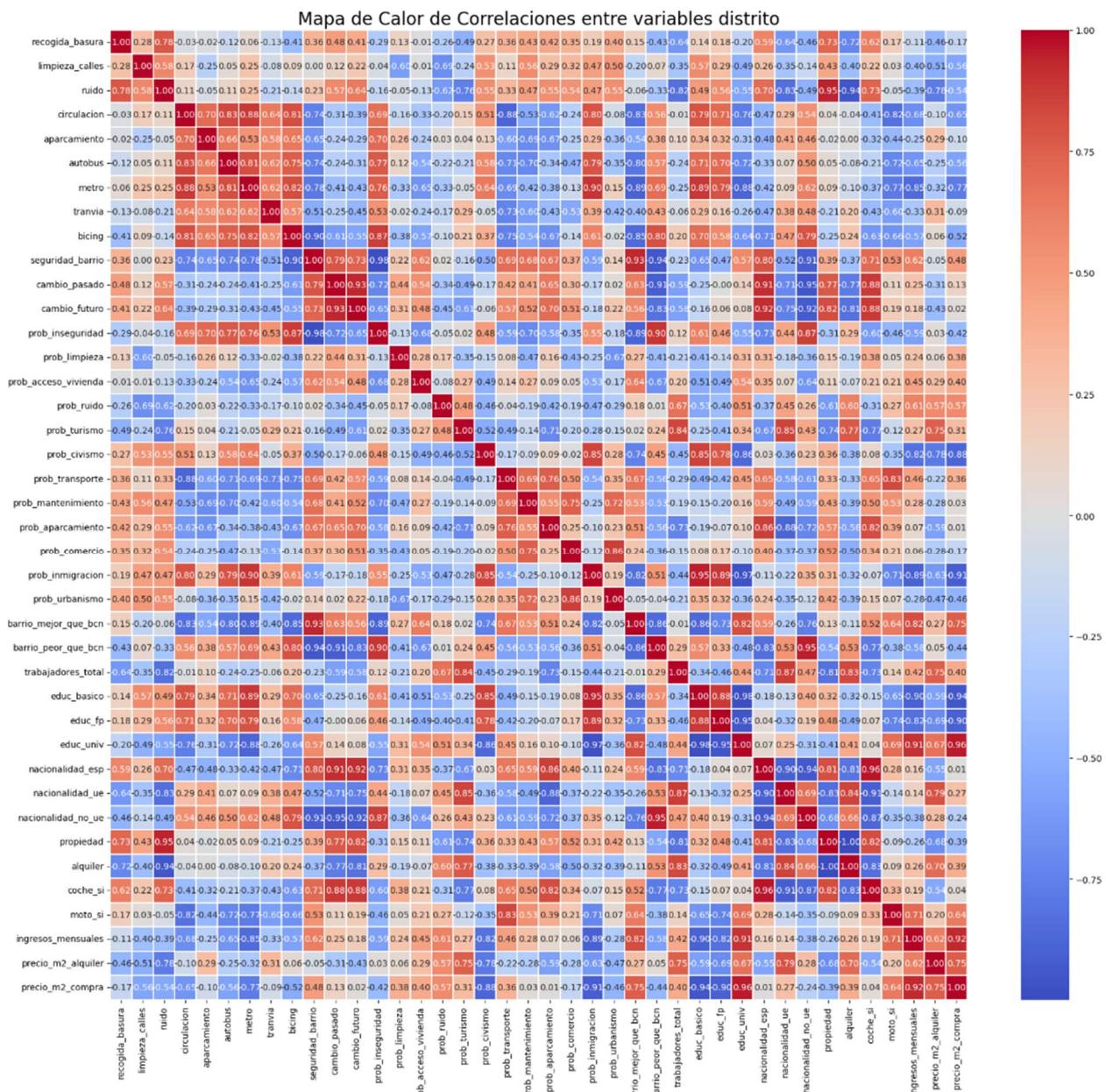
Tabla 14 Dataset df_distrito

Existen algunas variables que se han calculado para evitar colinealidad por ejemplo valoración de ha mejorado o ha empeorado el barrio. Para conservar el valor de los datos se ha decidido conservar el índice neto, es decir, la resta de las valoraciones positivas y las negativas. Obteniendo un valor que en caso de ser positivo indica tendencia de mejora y en caso de ser negativo indica tendencia de empeoramiento.

En cuanto a las variables de problema más grave del barrio, solo se han seleccionado las variables que tenían en algún barrio un porcentaje mayor o igual al 5 %. De esta manera se evita aportar más ruido a un dataset con muchas variables tratando de conservar las que puedan aportar valor. En este dataset se recogen las opiniones de los vecinos de los distritos, lo cual es una fuente de conocimiento muy potente. Sin embargo, siempre existe la duda de si los datos son transparentes o no, debido a que una encuesta promovida por un organismo público puede ser modificada para maquillar o esconder algunos datos y publicitar o enfatizar otros. Las variables de precio medio por m² de alquiler y compraventa se han obtenido de Idealista (52), si bien, también existe este dato en el portal de open data Bcn, se ha preferido escoger un dato externo para enriquecer las fuentes de datos.

Analisis de dataset

Se analiza la correlación de las variables del dataset creado a partir de la encuesta del Ajuntament de Barcelona y los datos obtenidos de precio medio de alquiler de Idealista.



pinta en rojo, si el valor de una variable aumenta y el otro disminuye se pinta en azul. Para determinar que es una correlación fuerte ya sea positiva o negativa se toma en consideración que el índice de correlación sea mayor a 0.8 o menor que -0.8 ya que al tener una gran cantidad de variables existen correlaciones que no aportan mucho valor como por ejemplo recogida de basura y ruido.

Las correlaciones más destacables son:

Variable ruido con variable vivienda en propiedad positiva 0.95 y con variable alquiler negativa -0.94. Esto nos indica que los propietarios de vivienda son los que tienen como problemática el ruido, además el alquiler al tener una relación negativa nos indica que los vecinos propietarios son más sensibles ante la problemática del ruido.

Existe relación positiva muy fuerte entre la valoración positiva de los usuarios de metro y problema de inmigración 0.9 y nivel de educación básica 0.89 y a su vez, negativa fuerte con la valoración de que el barrio es mejor que otros 0.89, educación universitaria 0.88 e ingresos mensuales 0.85. Esto nos muestra que los usuarios de metro se encuentran en distritos donde suelen tener menos ingresos y se tienen que desplazar a otros distritos que no son el suyo y que los encuestados que tienen titulación universitaria no son usuarios de metro o lo valoran negativamente.

Los usuarios del servicio bicing muestran correlación con los encuestados que su mayor problema es la inseguridad 0.87 y negativa muy fuerte con los que destacan la seguridad 0.90 esto puede deberse a que los usuarios de bicing no circulan por los distritos que son considerados más seguros o que circulan por distritos con alta población o zonas turísticas donde la sensación de seguridad es menor.

Los encuestados que más han destacado la sensación de seguridad en el barrio también consideran que su barrio es mejor que el resto 0.93 y tiene una correlación negativa con encuestados con nacionalidad no europea -0.91. Este es un indicativo importante ya que parece que los problemas de seguridad están relacionados con residentes no europeos y por tanto con la inmigración. Así se muestra también en los índices calculados de sensación de mejora del barrio tanto actual como al año siguiente ya que ambos muestran relación positiva muy fuerte con nacionalidad española y muy fuerte negativa con residentes con nacionalidad no europea y relación significativa positiva con la variable vivienda en propiedad 0.77 y 0.82 y negativa con alquiler -0.77, 0.81. También se ve reflejado en los que consideran que el mayor problema es la inseguridad y los residentes que no son de la unión europea 0.87 además piensan que su barrio es peor que el resto de los distritos 0.90. En cuanto al

problema de civismo se encuentra relación con problema de inmigración 0.85 y nivel de educación básica 0.85. los residentes no europeos y con el nivel de educación básica 0.95. Estas correlaciones no tienen por qué implicar causalidad ya que puede deberse a desigualdad económica o vulnerabilidad social (53).

De la correlación entre las variables se puede extraer que existe diferenciación de distritos en cuanto a transporte, renta, seguridad, calidad de vida y respecto al perfil de los residentes, mayor nivel de estudios, mayor renta y precios más elevados de vivienda. Presencia de extranjeros, nivel bajo de educación en el barrio denotan una sensación de inseguridad y precios de vivienda más bajos.

PCA

Cuando se disponen de tantas variables, es muy complejo y costoso realizar un trabajo de clasificación o predicción. Para poder realizar esta tarea se ha decidido por reducir el número de dimensiones mediante el análisis de componentes principales (PCA) (54) ya une y transforma las variables que están correlacionadas en nuevas variables reduciendo el número de variables. Este número reducido de variables se llama componentes principales y se debe de seleccionar el número de componentes principales que se desea obtener. Existen técnicas como Scree Plot (55) que marcan un punto en una gráfica que indica el número de componentes principales que explica mejor la varianza. Otra técnica es establecer la varianza explicada que se desea obtener así de esta manera, se ajusta el número de componentes principales a unos resultados mínimos esperados. En este caso se ha decidido que los componentes principales han de explicar el 90 % de la varianza y se ha obtenido que 5 componentes principales explican el 91.12% de la varianza. Una vez realizado el PCA se obtiene la tabla de pesos de las variables para cada componente principal. Se puede analizar la tabla, pero cuando existen muchas variables una buena solución es generar un mapa de calor para identificar de manera más visual las variables que más influyen en cada componente principal.

COMPONENTES	VARIABLES DESTACABLES	INTERPRETACIÓN
PC1	Positivamente (+): Tasa_C_Personas, Tasa_C_Patrimonio, bicing, prob_inseguridad. Negativamente (-): Seguridad_barrio, coche_si, nacionalidad_esp	PC1 — Intensidad urbana, movilidad e inseguridad Esta componente está marcada por las tasas de criminalidad en delitos contra personas (como agresiones, amenazas) y el patrimonio (robos, hurtos). También influye el transporte de bicing y la valoración de problema de inseguridad
PC2	Positivamente (+): Educ_universitaria, ingresos_mensuales, precio_m2_compra, alquiler Negativamente (-): Educ_fp, educ_basico, prob_civismo, prob_inmigracion	PC2 — Nivel económico y educativo Esta componente diferencia distritos con mayor nivel educativo, ingresos más elevados y precios de vivienda altos frente a zonas con menor nivel educativo y menor renta.
PC3	Positivamente (+): prob_limpieza, aparcamiento, autobus, tranvia, propiedad. Negativamente (-): prob_urbanismo, prob_mantenimiento, prob_comercio, autobús, alquiler	PC3 — Zona residencial y transporte Esta componente destaca la movilidad y el régimen de propiedad frente al alquiler. Zonas residenciales alejadas del centro.
PC4	Positivamente (+): Tasa_Trans_Publico, Tasa de movilidad por superficie, Tasa_Trans_privado, prob_turismo, prob_inmigracion. Negativamente: (-) prob_limpieza, prob_seguridad prob_transporte	PC4 — Conectividad de transporte público Esta componente refleja distritos con alta densidad de infraestructuras de transporte y movilidad, percepción negativa en limpieza y seguridad.
PC5	Positivamente (+): prob_comercio aparcamiento recogida_basura prob_ruido prob_urbanismo Negativamente (-): prob_acceso_vivienda limpieza_calles prob_aparcamiento	PC5 - Actividad comercial, núcleo urbano Esta componente destaca zonas con actividad comercial y problemáticas asociadas al urbanismo donde el acceso a vivienda no es percibido como un problema central. .

Tabla 15 Descripción componentes principales

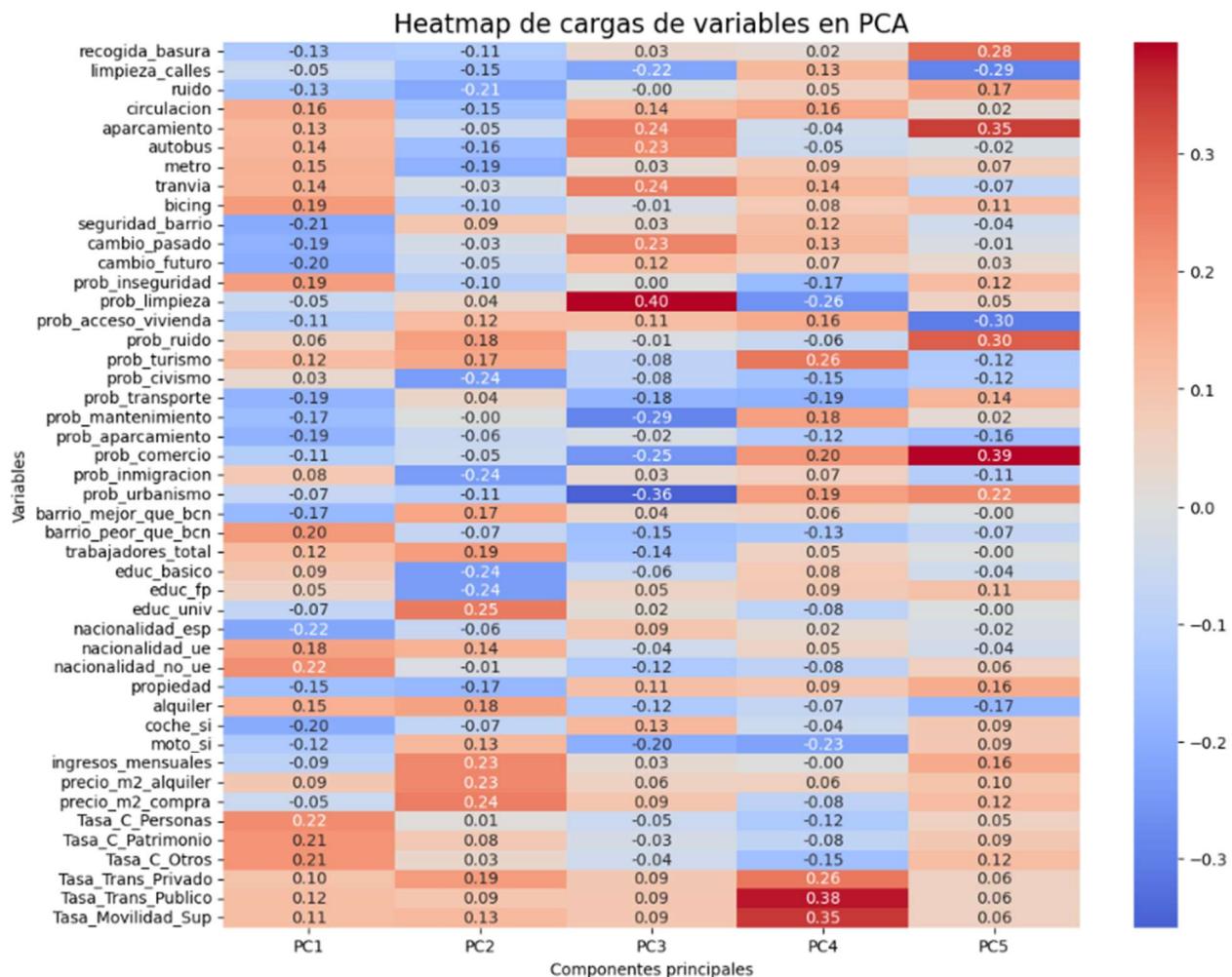


Figura 16 Mapa de Calor Carga de variables en PCA

Una vez conseguida la reducción de dimensionalidad, se puede trabajar con un algoritmo de agrupación (clustering), en este caso se ha decidido utilizar un algoritmo de aprendizaje no supervisado ya que no se dispone de un conjunto de datos etiquetado.

Kmeans

El algoritmo KMeans es un algoritmo que se basa en centroides, dividiendo un conjunto de datos en grupos similares en función de la distancia al centroide. Un valor de k alto implica que existen más centroides y por tanto se tiene más detalle, por el contrario, un valor de k más bajo implica que haya menos detalle y se puede perder información relevante.

Kmeans necesita que se le indique el valor k de antemano. Para obtener el valor k existen dos técnicas muy utilizadas el método de Elbow y el método de Silhouette. El método de Elbow mide

la distancia euclíadiana entre cada punto y el centro del clúster. No siempre tener más detalle es mejor ya que el costo computacional puede ser muy elevado.

Se ha de escoger el número de clústeres, para ello se va a realizar el método Elbow y silhouette

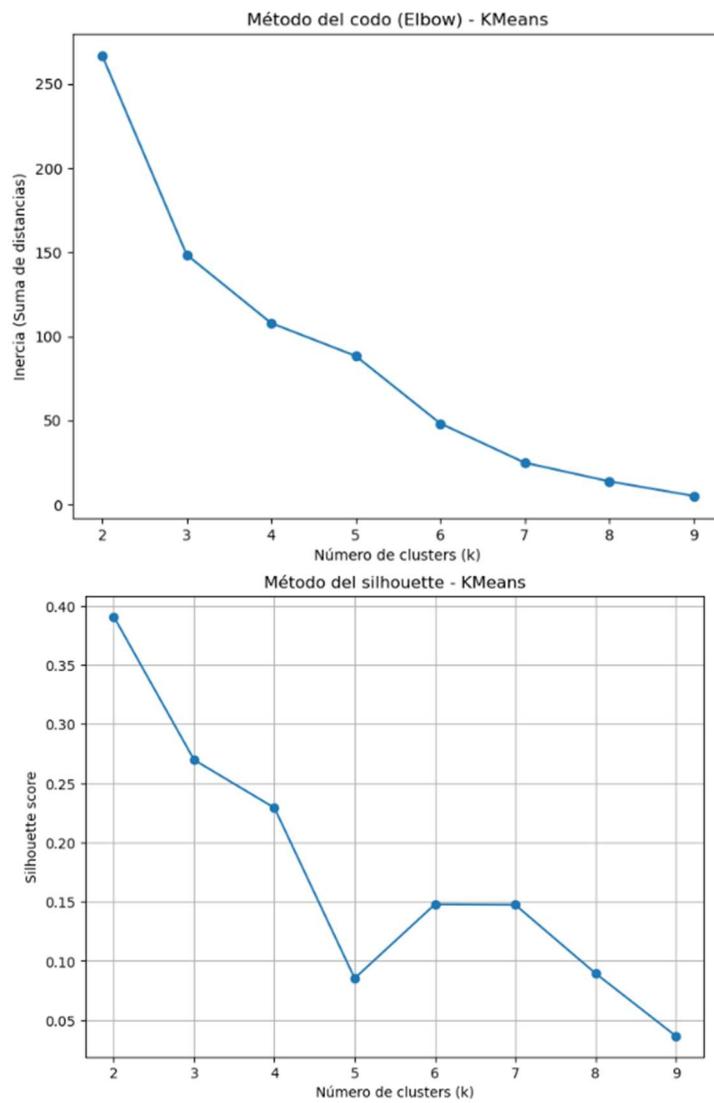


Figura 17 Resultados Métodos Elbow y Silhouette

Se observa en el método de Elbow como a partir de 3 la pendiente de la recta pierde inclinación por lo que un posible número válido para k es 3. Se observa con el método de Silhouette como el punto con el score más alto es el número 2. Sin embargo, el 3 también tiene buena puntuación y podremos tener más detalles en la clasificación de distritos. Por tanto, se selecciona como número óptimo de clúster 3. Una vez seleccionado el valor de k se procede a ejecutar Kmeans.

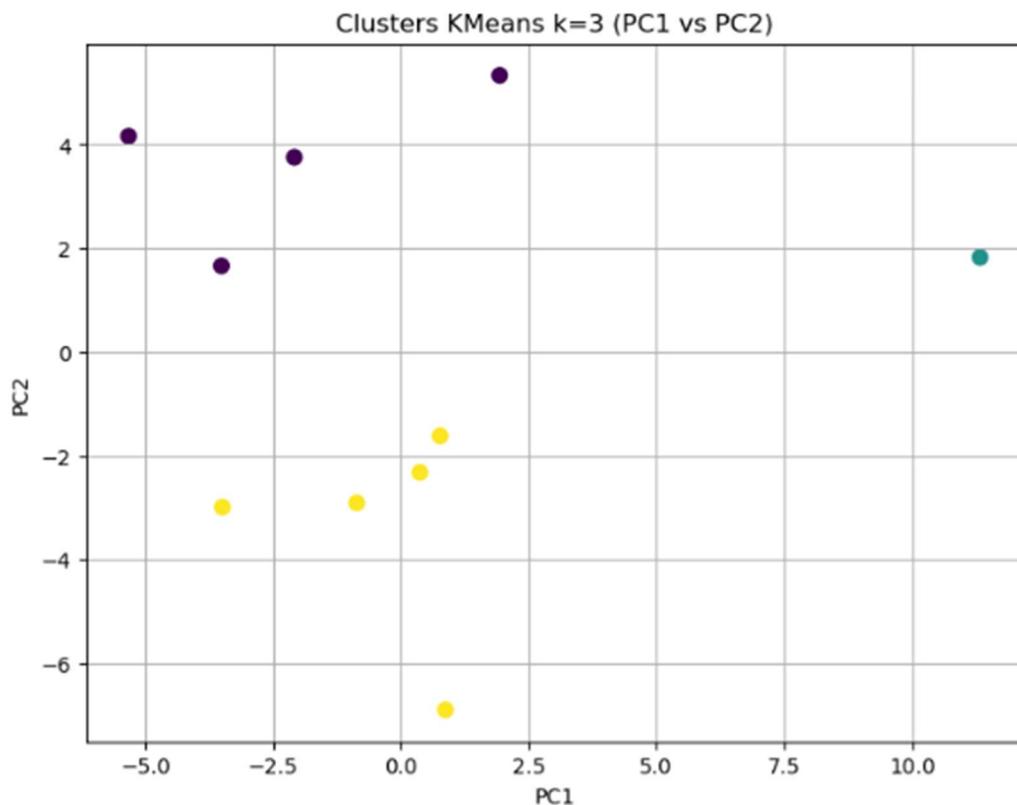


Figura 18 Clústers KMeans

Se observa como Kmeans ha agrupado los 10 distritos en 3 grupos. El grupo violeta en la parte superior izquierda el grupo amarillo en la parte inferior izquierda y un único distrito en el clúster pintado de verde. Con la descripción de los componentes principales que se ha realizado antes se puede interpretar que los puntos violetas que están más cerca de PC2 son aquellos distritos que tienen mejor nivel de estudios y mayor poder adquisitivo. En grupo amarillo más cerca del PC1 serían aquellos distritos que tienen problemas sociales.

El método Kmeans permite obtener una tabla con la relación entre las variables y clúster,

Se obtiene la tabla y se observa como:

- El clúster 0 tiene como distritos asignados el distrito de Les Corts, Gràcia, Sarrià-Sant Gervasi y Eixample.
- El clúster 1 tiene como distrito asignado Ciutat Vella
- El clúster 2 tiene como distritos asignados Nou Barris, Sant Martí, Horta-Guinardó, Sants-Montjuïc, Sant Andreu.

Cluster	0	1	2
recogida_basura	7.22	6.80	7.32
limpieza_calles	6.60	6.60	6.74
ruido	5.85	5.50	6.16
circulacion	5.10	5.60	5.50
aparcamiento	3.92	4.40	4.20
autobus	7.15	7.50	7.42
metro	7.20	7.60	7.48
tranvia	7.60	7.90	7.62
bicing	6.15	6.90	6.46
seguridad_barrio	6.60	4.80	5.98
cambio_pasado	10.15	-28.20	9.02
cambio_futuro	31.65	8.20	33.72
prob_inseguridad	12.65	45.60	24.46
prob_limpieza	9.35	6.60	8.84
prob_acceso_vivienda	9.30	3.20	6.94
prob_ruido	6.80	6.90	4.94
prob_turismo	6.32	11.20	1.40
prob_civismo	2.42	3.90	4.80
prob_transporte	4.47	0.30	3.04
prob_mantenimiento	3.45	0.30	2.88
prob_aparcamiento	3.28	0.00	2.94
prob_comercio	2.12	1.00	2.66
prob_inmigracion	0.80	2.70	3.34
prob_urbanismo	1.27	1.00	2.18
barrio_mejor_que_bcn	84.48	38.40	58.24
barrio_peor_que_bcn	1.40	30.50	8.00
trabajadores_total	56.65	61.10	53.90
educBasico	35.90	47.20	48.66
educ_fp	10.77	14.60	19.46
educ_univ	51.45	36.00	28.84
nacionalidad_esp	77.92	48.00	78.04
nacionalidad_ue	9.30	19.80	6.80
nacionalidad_no_ue	12.80	32.10	15.14
propiedad	48.55	28.00	59.48
alquiler	48.02	68.40	36.82
coche_si	46.55	19.30	47.78
moto_si	20.67	15.80	15.76
ingresos_mensuales	3088.40	2477.30	2465.44
precio_m2_alquiler	23.90	26.30	18.92
precio_m2_compra	6140.50	4796.00	3967.20
Tasa_C_Personas	4.70	18.08	6.37
Tasa_C_Patrimonio	59.86	247.22	47.09
Tasa_C_Otros	3.80	15.57	4.30
Tasa_Trans_Privado	15.05	18.07	4.28
Tasa_Trans_Publico	57.49	68.49	44.86
Tasa_Movilidad_Sup	72.57	86.80	49.18

Tabla 16 Relación variables con los clústers

Se observa como el clúster 0 tiene como variables significativas barrio_mejor_que_bcn, precio de m² de compra más alto, los ingresos mensuales más altos, el nivel de estudios más alto y son los que menos consideran que la inseguridad sea un problema. Por tanto, se puede resumir en que el clúster 0 son distritos de clase alta, acomodados.

Se observa que el clúster 1 tiene el peor índice que cambio_pasado es decir considera que el distrito ha empeorado, es el que más valora el problema de inseguridad, el que más trabajadores en activo tiene y los que más residen de alquiler. También se observa la Tasa de criminalidad contra el patrimonio más elevada, así como la Tasa de criminalidad contra las personas. También valoran que su distrito es peor que el resto de Barcelona. Por tanto, se puede resumir que el clúster 1 clasifica distritos conflictivos de clase obrera, con poca expectativa en el distrito

Se observa que el clúster 2 tiene como variables significativas la educación básica, son los distritos que menor precio por m² cuadrado tienen tanto en alquiler como en compraventa, suelen tener coche y la vivienda en propiedad. Consideran que el urbanismo y el comercio son un problema. Son también los que más expectativas tienen de cambio futuro en el distrito. Por tanto, se puede decir que son distritos de clase media baja, con poca renta y nivel de estudios.

Se ha podido mediante las estadísticas de una encuesta, datos oficiales abiertos y el cálculo de tasas clasificar los distritos de Barcelona según sus características sociodemográficas con Análisis de Componentes Principales y el algoritmo de aprendizaje no supervisado Kmeans/

3.2 Análisis de mercado y predicción de precios de viviendas

En la segunda parte del trabajo se realiza un análisis del mercado y predicción del valor de las viviendas. Los datos se obtienen de Idealista mediante el acceso a su API. Para conseguir el acceso es necesario solicitar permiso (56), tras aceptar la solicitud se recibe correo con la APIkey y Secret, la documentación informativa de solicitud de token de acceso y manual de consultas de la API. Las consultas en la API de Idealista están limitadas a 100 cada mes, y cada consulta puede albergar como máximo 50 registros por tanto se tiene una limitación de 5000 registros al mes.

Debido a la limitación se ha decidido realizar consultas en 3 niveles diferentes para poder llevar a cabo el trabajo final de máster, consultas de compraventa, consultas de alquiler y consultas de habitaciones. De esta manera se puede realizar una predicción de precios de vivienda, y se puede realizar sugerencias al usuario de la visualización en caso de estar interesado en alquilar una habitación o un inmueble en determinado distrito.

La distribución de las consultas queda de la siguiente manera.

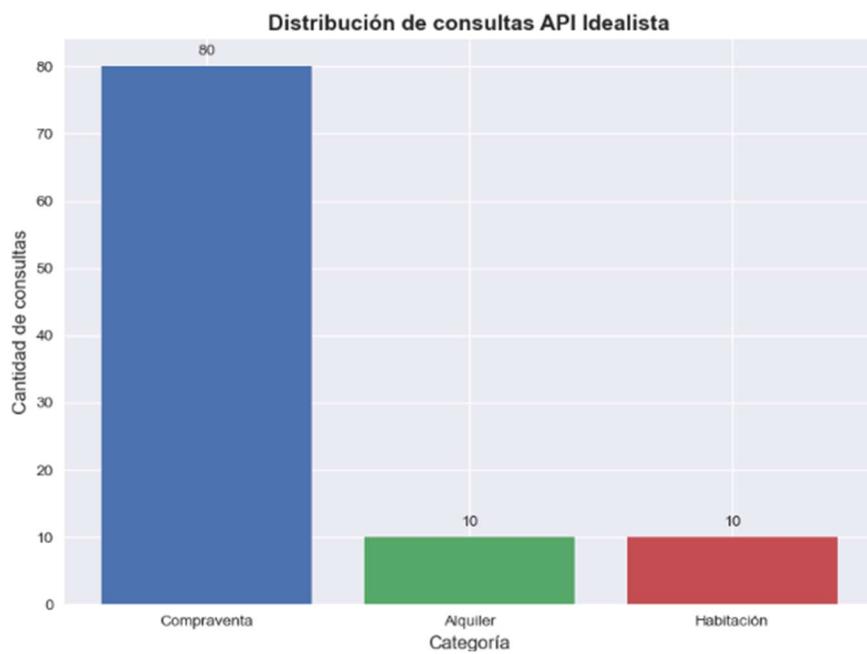


Figura 19 Distribución Consultas API Idealista

Con la distribución escogida se pueden obtener 4000 registros ($80 * 50$) de viviendas en compraventa, 500 registros ($10 * 50$) de viviendas en alquiler y 500 registros ($10 * 50$) de habitaciones en régimen de alquiler.

Una vez revisada la documentación de consultas se seleccionan los parámetros disponibles para realizar filtrado de datos para cada uno de los conjuntos de datos con el objetivo de obtener información relevante y evitar valores extremos (outliers) que puedan distorsionar el análisis posterior.

Como elementos de búsqueda común a los 3 conjuntos de datos se aplica:

Fecha de publicación: último mes.

Ordenación: anuncios más recientes primero.

Estos criterios permiten obtener conjunto de datos actualizados

- Para el conjunto de alquiler de habitaciones se aplican los siguientes filtros:

Ubicación: ciudad de Barcelona.

Precio mínimo: 300 €

Objetivo excluir habitaciones de corta estancia, alquiler turístico o temporal.

Precio máximo: 900 €

Valores superiores permiten alquilar una vivienda completa.

- Para el conjunto de alquiler de vivienda, se aplican los siguientes filtros:

Ubicación: ciudad de Barcelona.

Precio mínimo: 600 €

Se descartan viviendas con precios bajos, estancias cortas o alquiler turístico.

Precio máximo: 4.000 €

Se trata de evitar la inclusión de alquileres de lujo, reduciendo la presencia de outliers.

El objetivo es obtener una distribución de precios más homogénea para análisis descriptivos y visualizaciones.

- Para el conjunto de compraventa de vivienda, se establecieron los siguientes filtros:

Ubicación: ciudad de Barcelona.

Precio mínimo: 80.000 €

Se trata de excluir viviendas ocupadas, en mal estado o con condiciones especiales.

Precio máximo: 1.000.000 €

El objetivo es obtener viviendas de búsqueda habitual excluyendo viviendas de lujo.

Para el acceso al token se trató de realizar mediante la librería requests, obteniendo en todos los intentos el error http 406 Not acceptable (57). Se probaron diferentes soluciones siendo la definitiva obtener el token mediante la herramienta curl (58) Una vez obtenido el token de acceso, se realizaron las consultas con curl en Python mediante subprocess.

Para la primera consulta, se realizó la selección de la primera página de habitaciones en alquiler para comprobar su funcionamiento y no malgastar peticiones ya que el límite son 100 al mes. Una vez obtenido satisfactoriamente los datos de la primera página se pasa a ejecutar consultas para habitaciones en alquiler, viviendas en alquiler y viviendas en compraventa con los filtros mencionados anteriormente obteniendo 3 datasets:

- barcelona_habitaciones.csv: Registros de alquiler de habitaciones
- barcelona_alquiler.csv: registros de viviendas en alquiler
- barcelona_compraventa.csv: Registros de viviendas en compraventa.

Una vez concluidas las consultas con los parámetros fijados, se han consumido 74 consultas quedando disponibles 26 para completar datos en caso de que tras la limpieza de datos sea necesario o para mejorar las visualizaciones realizando consultas para completar viviendas en ciertos distritos.

3.2.1 Análisis exploratorio

Dataset barcelona_habitaciones.csv

El dataset barcelona_habitaciones contiene el registro de las 500 habitaciones en alquiler más recientes publicadas en Idealista en el municipio de Barcelona en el momento de la extracción. Fecha de extracción 21/12/25

Nombre dataset	Número de filas	Número de columnas
Barcelona_habitaciones	500	43
Nombre columna	Valores únicos	Descripción
Data columns (total 43 columns): # Column Non-Null Count Dtype 0 propertyCode 499 non-null int64 1 thumbnail 493 non-null object 2 numPhotos 499 non-null int64 3 floor 414 non-null object 4 price 499 non-null float64 5 priceInfo 499 non-null object 6 propertyType 499 non-null object 7 operation 499 non-null object 8 size 499 non-null float64 9 rooms 499 non-null int64 10 bathrooms 499 non-null int64 11 address 499 non-null object 12 province 499 non-null object 13 municipality 499 non-null object 14 district 499 non-null object 15 country 499 non-null object 16 neighborhood 499 non-null object 17 latitude 499 non-null float64 18 longitude 499 non-null float64 19 showAddress 499 non-null bool 20 url 499 non-null object 21 description 467 non-null object 22 hasVideo 499 non-null bool 23 newDevelopment 499 non-null bool 24 tenantNumber 499 non-null int64 25 hasLift 493 non-null object 26 isSmokingAllowed 499 non-null bool 27 priceByArea 499 non-null float64 28 change 499 non-null object 29 detailedType 499 non-null object 30 suggestedTexts 499 non-null object 31 hasPlan 499 non-null bool 32 has3DTour 499 non-null bool 33 has360 499 non-null bool 34 hasStaging 499 non-null bool 35 isOnlineBookingActive 499 non-null bool 36 savedAd 499 non-null object 37 notes 499 non-null object 38 topNewDevelopment 499 non-null bool 39 newDevelopmentHighlight 499 non-null bool 40 topPlus 499 non-null bool 41 tenantGender 414 non-null object 42 externalReference 163 non-null object	499	Se obtienen 499 registros únicos de 500 registros totales. Se observa como existen variables con datos nulos. externalReference 336 tenantGender 85 floor 85 description 32 hasLift 6 thumbnail 6. El tipo de datos es variado existen columnas con valores int64, float64, booleanos, y object (cadenas de texto.)

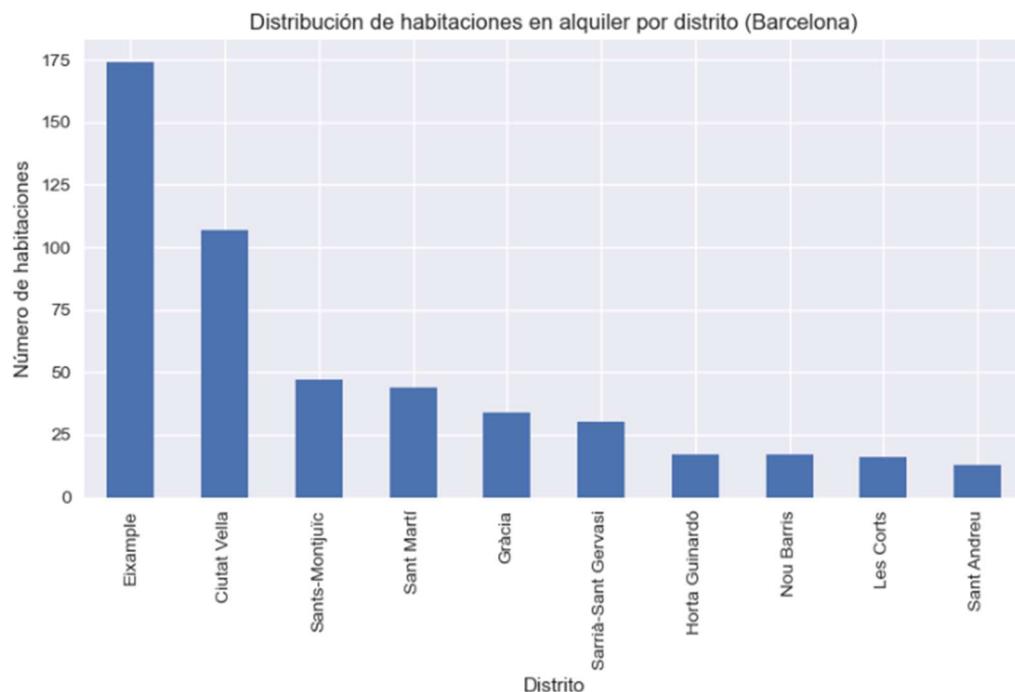
Acciones llevadas a cabo

Se decide eliminar columna externalReference ya que no aporta información relevante. Se decide conservar la columna thumbnail ya que solo presenta 6 valores nulos y puede aportar para la visualización. Para los registros nulos, no se realiza ninguna acción. De igual manera para description,hasLift, floor y tenantGender ya que para la columna floor y hasLift en este caso no se va a realizar una regresión con los datos de habitaciones en alquiler y el genero no es relevante para el estudio.

Tabla 17 Dataset barcelona_habitaciones

Este dataset se utilizará para analizar el mercado de habitaciones en alquiler y para ofrecer opciones en la visualización de viviendas según el clúster y barrio asignado.

Se puede observar en la distribución de habitaciones en alquiler por distrito que existen distritos con más oferta de habitaciones en alquiler que otros. Por ejemplo, Eixample y Ciutat Vella destacan por la oferta de habitaciones en alquiler respecto al resto.

*Figura 20 Distribución de habitaciones en alquiler por distrito*

Se calcula el precio medio de habitación en alquiler el cual es de 615.13 euros. Se observa en la distribución como existen distritos que tienen un precio medio más alto Sarriá-Sant Gervasi cerca de 700 euros frente a Sant Andreu y Nou Barris que rondan los 500 euros.

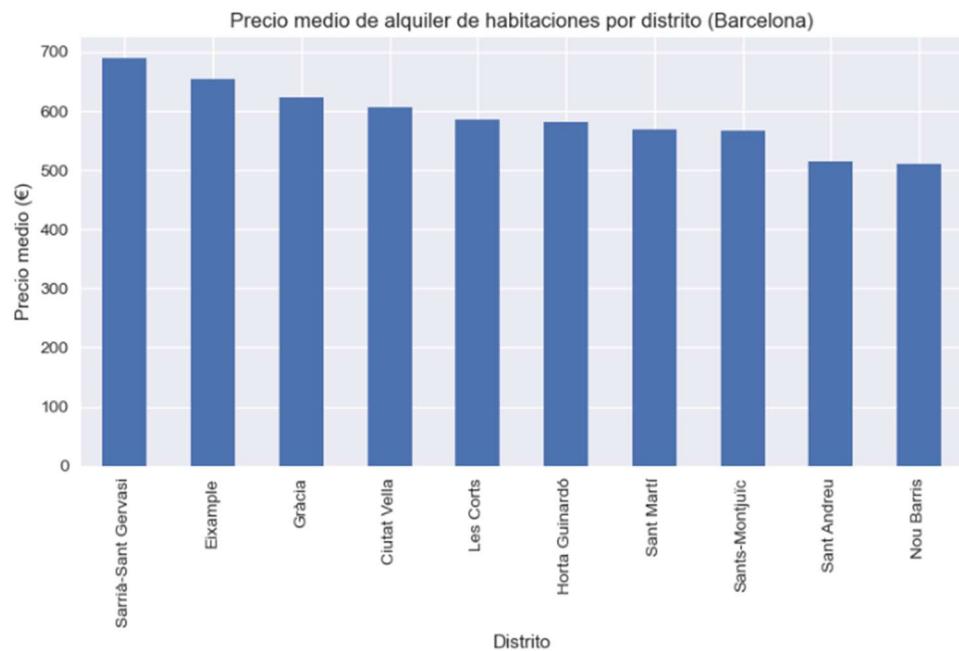


Figura 21 Precio medio de alquiler de habitaciones por distrito

Dataset barcelona_alquiler.csv

El dataset barcelona_alquiler contiene el registro de las 500 viviendas en alquiler más recientes publicadas en Idealista en el municipio de Barcelona en el momento de la extracción. Fecha de extracción 21/12/25

Nombre dataset			Número de filas	Número de columnas
Barcelona_alquiler			500	42
Nombre columna			Valores únicos	Descripción
Column	Non-Null Count	Dtype	500	
0 propertyCode	500 non-null	int64		
1 thumbnail	499 non-null	object		
2 externalReference	380 non-null	object		
3 numPhotos	500 non-null	int64		
4 floor	455 non-null	object		
5 price	500 non-null	float64		
6 priceInfo	500 non-null	object		
7 propertyType	500 non-null	object		
8 operation	500 non-null	object		
9 size	500 non-null	float64		
10 rooms	500 non-null	int64		
11 bathrooms	500 non-null	int64		
12 address	500 non-null	object		
13 province	500 non-null	object		
14 municipality	500 non-null	object		
15 district	500 non-null	object		
16 country	500 non-null	object		
17 neighborhood	500 non-null	object		

18 latitude	500 non-null	float64	
19 longitude	500 non-null	float64	
20 showAddress	500 non-null	bool	
21 url	500 non-null	object	
22 description	500 non-null	object	
23 hasVideo	500 non-null	bool	
24 status	500 non-null	object	
25 newDevelopment	500 non-null	bool	
26 hasLift	490 non-null	object	
27 priceByArea	500 non-null	float64	
28 change	500 non-null	object	
29 detailedType	500 non-null	object	
30 suggestedTexts	500 non-null	object	
31 hasPlan	500 non-null	bool	
32 has3DTour	500 non-null	bool	
33 has360	500 non-null	bool	
34 hasStaging	500 non-null	bool	
35 savedAd	500 non-null	object	
36 notes	500 non-null	object	
37 topNewDevelopment	500 non-null	bool	
38 newDevelopmentHighlight	500 non-null	bool	
39 topPlus	500 non-null	bool	
40 exterior	455 non-null	object	
41 parkingSpace	36 non-null	object	

Acciones llevadas a cabo			
Se decide eliminar columna externalReference ya que no aporta información relevante. Se realiza conteo de los valores existentes de la variable parkingSpace y al parecer los valores NaN corresponden con las viviendas que no disponen de plaza de parking. Se asume que esto es así y se recodifican los valores de la variable de Nan a false			

Tabla 18 Dataset barcelona_alquiler

Este dataset se utilizará para analizar el mercado de viviendas en alquiler y para ofrecer opciones en la visualización de viviendas según el clúster y barrio asignado. Se puede observar en la distribución de viviendas en alquiler por distrito que existen distritos con más oferta de viviendas en alquiler que otros tal y como sucedía en el dataset de habitaciones.

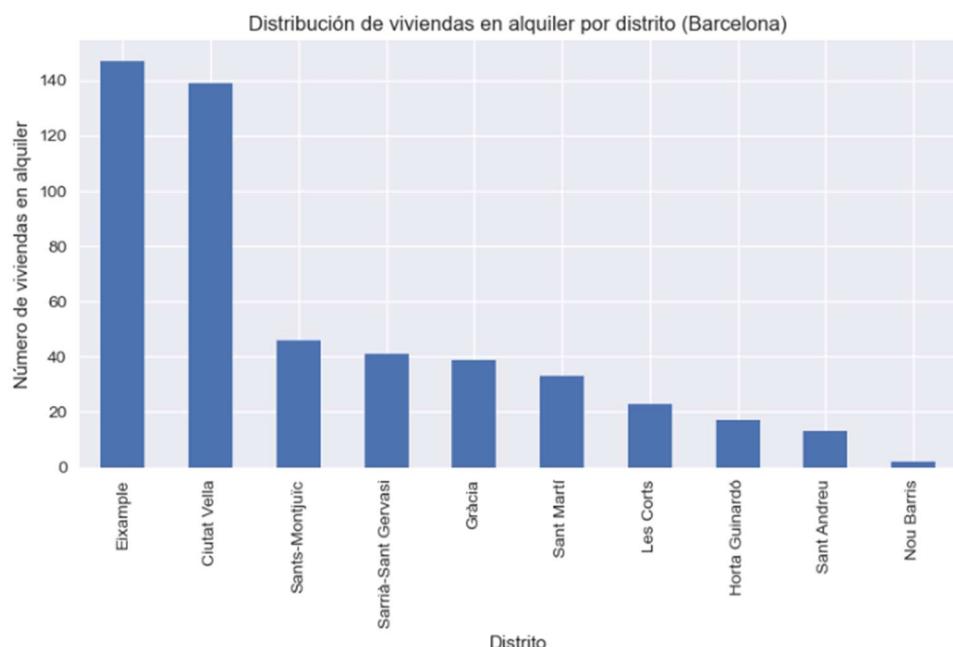


Figura 22 Distribución de viviendas en alquiler por distrito

Se calcula el precio medio de vivienda en alquiler el cual es de 1792.22 euros. Se observa en la distribución como existen distritos que tienen un precio medio más alto Les Corts, Sant Martí, Eixample y Sarriá-Sant Gervasi alrededor de 2000 euros frente a Sant Andreu y Nou Barris que rondan los 1300 euros. Si bien en este caso puede no deberse solo al barrio sino a las características de la vivienda como habitaciones, metros cuadrados, amenities, estado de la vivienda etc.

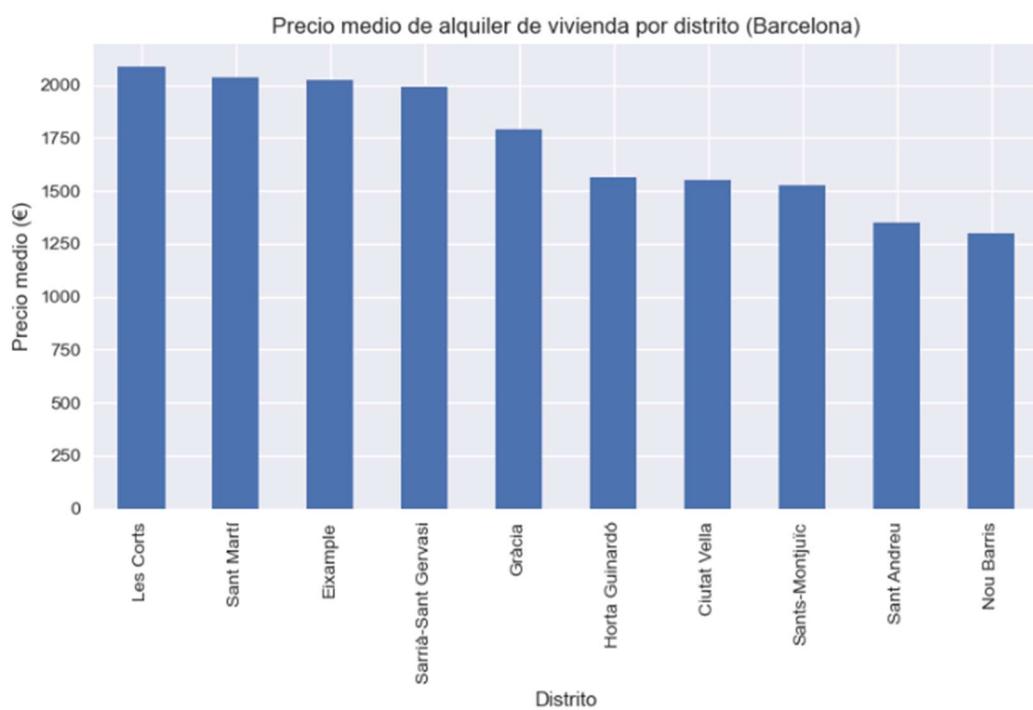


Figura 23 Precio medio de viviendas en alquiler por distrito

En cuanto a la distribución de precios de los registros se obtiene lo siguiente:

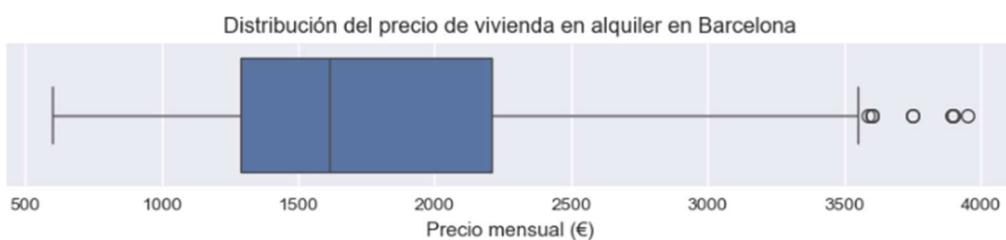


Figura 24 Distribución de precio de viviendas en alquiler por distrito

El boxplot permite conocer de manera muy visual la distribución de los datos. La caja de color azul representa rango intercuartílico (IQR) de izquierda a derecha el percentil 25(Q1) el percentil 50 o mediana (Q2, línea dentro de la caja) y el percentil 75 (Q3) que indican el

número de viviendas con precios que están por debajo del percentil. El rango intercuartílico (la caja azul) representa el 50 por ciento central de los precios. Las líneas rectas a la izquierda y derecha que finalizan con una línea perpendicular (bigotes) indican el resto de los precios que no son considerados valores atípicos es decir se encuentran en el rango de 1.5 veces el IQR ($1.5 * \text{IQR}$). Los círculos exteriores son los valores atípicos (outliers). Analizando el boxplot podemos ver como el 75 por ciento de los precios de viviendas de alquiler están por debajo de los 2300 euros. Y la mediana se sitúa en torno a los 1700 (tal y como hemos calculado anteriormente). Los outliers por encima de 3500 pueden representar viviendas de lujo o muy grandes en zonas de lujo.

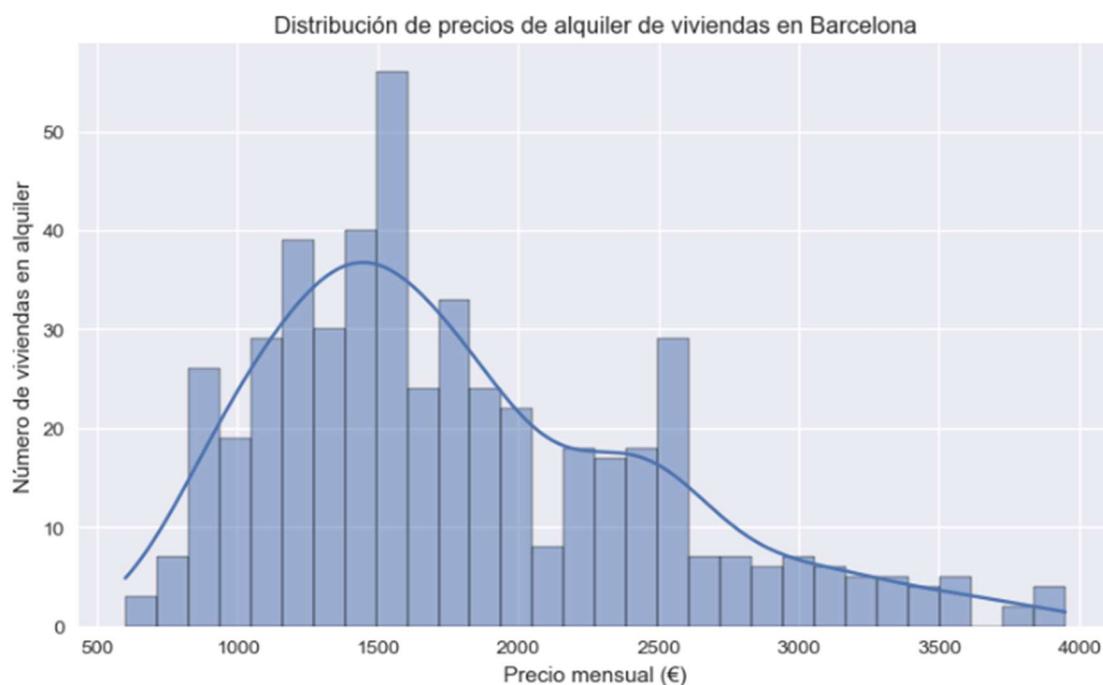


Figura 25 Distribución de precio de viviendas en alquiler por distrito

Se observa en el gráfico de distribución como el mayor número de viviendas se encuentra entre 1500 y 1600 euros. Se observa como la mayoría de las viviendas se encuentra entre los 1300 y 2000 euros y pasados los 2000 euros comienza a descender el número de viviendas. Esta forma alargada se llama asimetría positiva (cola a la derecha).

Se observa que entre los 2300 y 2600 euros hay un repunte de viviendas, puede deberse a distritos más caros o viviendas más grandes. A pesar de haber seleccionado los datos con el objetivo de filtrar outliers, existen algunos en el dataset de vivienda de alquiler.

Dataset barcelona_compraventa.csv

El dataset barcelona_compraventa contiene el registro de las 2607 viviendas en alquiler más recientes publicadas en Idealista en el municipio de Barcelona en el momento de la extracción. Fecha de extracción 21/12/25

Nombre dataset		Número de filas		Número de columnas
barcelona_compraventa		2607		44
Nombre columna		Valores únicos	Descripción	
Column	Non-Null Count	Dtype	500	Se observan valores nulos en varias columnas los cuales se han de procesar para realizar las técnicas de machine learning. El tratamiento de valores nulos va a ser tratado mediante eliminación, imputación, recodificación según el tipo de variable y la influencia que puede tener en el trabajo. Se realiza estudio de las variables en el código y se redacta la justificación en la memoria.
0 propertyCode	2607 non-null	int64		
1 thumbnail	2598 non-null	object		
2 externalReference	2122 non-null	object		
3 numPhotos	2607 non-null	int64		
4 floor	2411 non-null	object		
5 price	2607 non-null	float64		
6 priceInfo	2607 non-null	object		
7 propertyType	2607 non-null	object		
8 operation	2607 non-null	object		
9 size	2607 non-null	float64		
10 rooms	2607 non-null	int64		
11 bathrooms	2607 non-null	int64		
12 address	2607 non-null	object		
13 province	2607 non-null	object		
14 municipality	2607 non-null	object		
15 district	2607 non-null	object		
16 country	2607 non-null	object		
17 neighborhood	2607 non-null	object		
18 latitude	2607 non-null	float64		
19 longitude	2607 non-null	float64		
20 showAddress	2607 non-null	bool		
21 url	2607 non-null	object		
22 description	2605 non-null	object		
23 hasVideo	2607 non-null	bool		
24 status	2598 non-null	object		
25 newDevelopment	2607 non-null	bool		
26 hasLift	2534 non-null	object		
27 priceByArea	2607 non-null	float64		
28 change	2607 non-null	object		
29 detailedType	2607 non-null	object		
30 suggestedTexts	2607 non-null	object		
31 hasPlan	2607 non-null	bool		
32 has3DTour	2607 non-null	bool		
33 has360	2607 non-null	bool		
34 hasStaging	2607 non-null	bool		
35 savedAd	2607 non-null	object		
36 notes	2607 non-null	object		
37 topNewDevelopment	2607 non-null	bool		
38 newDevelopmentHighlight	2607 non-null	bool		
39 topPlus	2607 non-null	bool		
40 exterior	2379 non-null	object		
41 highlight	1033 non-null	object		
42 parkingSpace	255 non-null	object		
43 newDevelopmentFinished	62 non-null	object		
Acciones llevadas a cabo				
Se decide eliminar columna externalReference, highlight y newDevelopmentFinished ya que no aportan información relevante. Se realiza conteo de los valores existentes de la variable parkingSpace y al parecer los valores NaN corresponden con las viviendas que no disponen de plaza de parking. Se asume que esto es así y se recodifican los valores de la variable de Nan a false				

Tabla 19 Dataset barcelona_compraventa

Se observa como el dataset tiene 2 valores nulos en description, 9 en thumbnail. Este caso no se realiza ninguna acción. No son variables que vayan a afectar al desarrollo. Existen 9 valores faltantes en status se va a considerar que el estado de la vivienda es bueno.

La variable hasLift tiene 73 valores NaN, como por lo general el disponer de ascensor es un valor añadido el hecho de no especificarlo puede indicar que no lo tenga. Los valores NaN se recodifican como False. La variable floor tiene 196 valores faltantes. Se va a realizar la recodificación de los NaN de exterior como desconocido. Se ha analizado el precio medio entre ser de exterior y no serlo y es muy significativo por lo que se opta por calificarlos como desconocido. La variable externalReference se elimina ya que no aporta nada al trabajo. La variable highlight contiene 1574 valores nulos. No representa una característica de la vivienda sino el tipo de anuncio en el portal. Se decide eliminarla. La variable parkingSpace se va a tratar como en el dataset de barcelona_alquiler. Se va a codificar los valores NaN por False. La variable newDevelopmentFinished indica si la vivienda de obra nueva está acabada. solo existen 2 registros que esté acabada la vivienda por tanto no va a aportar un gran valor al estudio. Se decide eliminar la variable. Se realiza análisis de la distribución de las viviendas por distrito. Se observa que Eixample y Ciutat Vella son los distritos que más viviendas publicadas tienen en el conjunto de datos seleccionado y Les Corts es el distrito con menos vivienda en compraventa en este dataset.

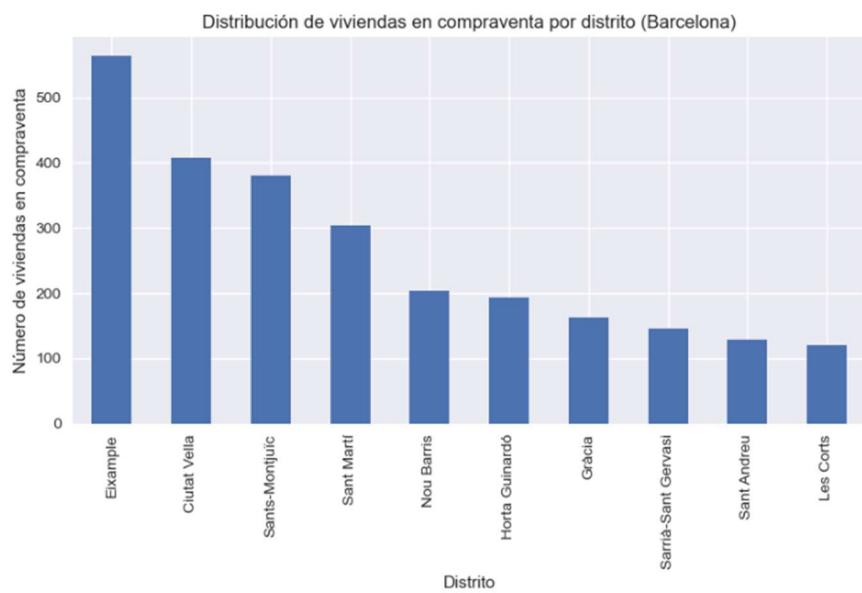


Figura 26 Distribución de viviendas en compraventa por distrito

Se calcula el precio medio de vivienda en compraventa el cual es de 446815.44 euros. Se observa en la distribución como existen distritos que tienen un precio medio más alto. Sarriá-Sant Gervasi alrededor de 700000 euros frente a Nou Barris que rondan los 250000 euros.

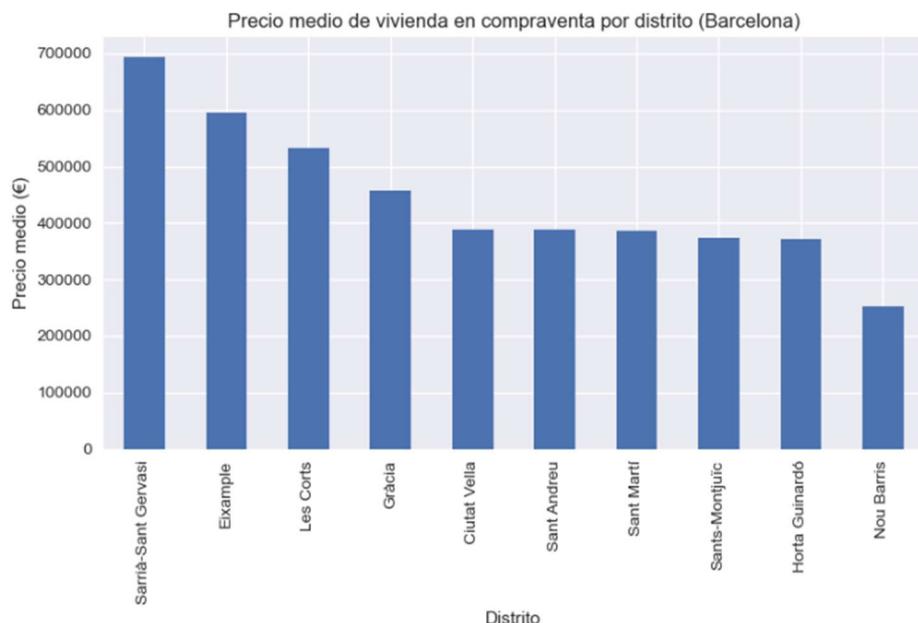


Figura 27 Precio medio de viviendas en compraventa por distrito

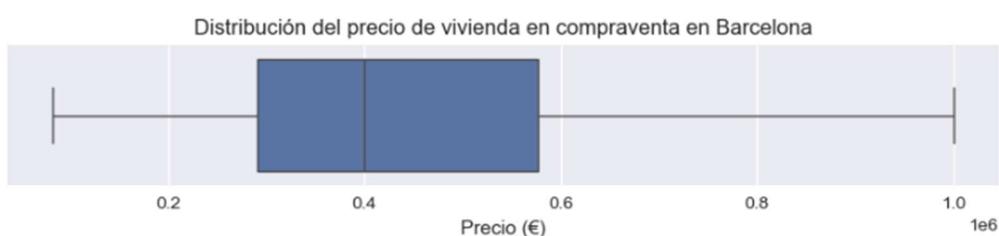


Figura 28 Distribución de precio de viviendas en compraventa por distrito

Como se puede ver en el Boxplot, no existen outliers. Eso quiere decir que hay suficientes datos en todo el conjunto de datos como para que la diferencia de precio no se considere como outlier. Los precios aparecen en millones de euros. Se observa que la mediana parece coincidir con los 400000 euros.

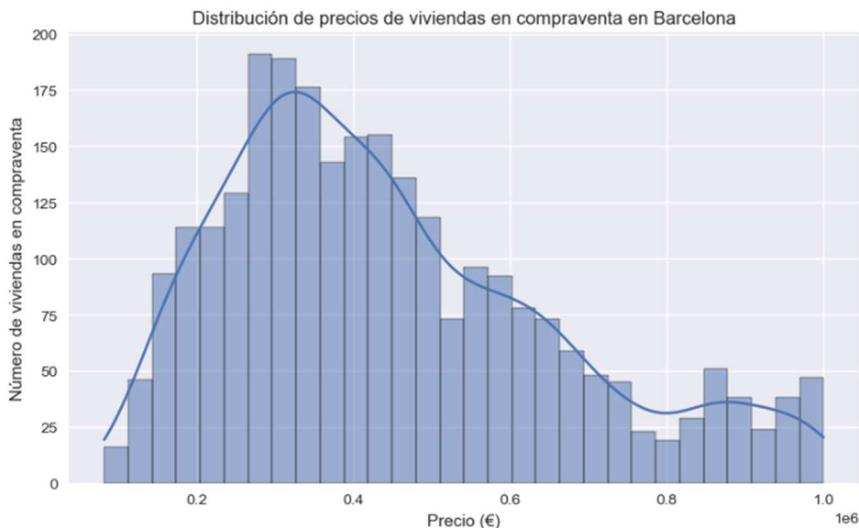


Figura 29 Distribución de precio de viviendas en compraventa

El conjunto de datos tiene asimetría positiva como el dataset barcelona_arquiler. La mayor parte de las viviendas extraídas se encuentran alrededor de 40000 euros.

A continuación, se prepara el conjunto de datos para realizar la predicción de precios. Para ello existen variables que son estadísticas o contienen información del anuncio de Idealista y que no aportan valor a la vivienda como número de fotos, o si el anuncio esta remarcado. También se eliminan variables que ocasionen fuga de datos (data leakage) (59) como precio por área y la dirección debido a su alta cardinalidad.

Análisis de relevancia mediante regularización (LASSO)

Para analizar la importancia de las variables de forma interpretable, se utilizó un modelo lineal con regularización L1 (LASSO)(60) permite identificar predictores relevantes al penalizar coeficientes y forzar a cero aquellos con contribución baja, facilitando la selección de variables. Las variables categóricas se incorporaron mediante codificación one-hot.

Se añadió al dataset de viviendas información a nivel de distrito, incorporando indicadores de calidad urbana, transporte, criminalidad y nivel socioeconómico. Para ello, se realizó una unión de los datos mediante un merge utilizando el nombre del distrito como clave. Se corrigió el preprocesado eliminando variables propias del anuncio (metadatos del portal, multimedia, textos y enlaces) y variables derivadas del objetivo (priceByArea) para evitar fuga de información y se añadieron los componentes principales obtenidos previamente. Cada registro de vivienda obtiene los valores del PCA de su distrito.

Modelos de predicción de precio de vivienda

Para la predicción del precio de vivienda se utilizan diferentes modelos cuyos resultados se recogen en un dataset para poder evaluar el rendimiento y decidir cuál es mejor para el estudio que se realiza. Para evitar sobreajuste y reducir el coste computacional, se empleó early stopping en aquellos modelos que lo permiten. El entrenamiento se detiene automáticamente cuando la métrica de validación no mejora tras un número determinado de iteraciones. Dado que el problema presenta relaciones no lineales, se priorizan modelos basados en árboles ya que permiten detectar las relaciones entre las variables categóricas y numéricas. Se incluyen también modelo lineal como línea base con el objetivo de evaluar la mejora de rendimiento que se consigue con estos modelos.

Además de los modelos evaluados, se consideraron otros enfoques habituales en problemas de regresión, como los métodos basados en distancia (KNN), las máquinas de soporte vectorial (SVM). Estos modelos fueron descartados debido a la sensibilidad a la escala de las variables y mayor coste computacional, así como los modelos basados en redes neuronales que requieren de volúmenes de datos mayores.

Como medidas de control se utiliza la validación cruzada, se emplea para estimar el rendimiento global y la capacidad de generalización del modelo, mientras que el análisis por tramos de precio se realiza sobre un modelo entrenado con un mayor volumen de datos, con el objetivo de estudiar el comportamiento del error en distintos segmentos del mercado

Modelo Ridge

El modelo Ridge se utiliza como una línea base para la predicción del precio de la vivienda. Permite controlar el sobreajuste cuando existe gran número de variables explicativas (61)

Modelo Random Forest

El modelo Random Forest está basado en árboles de decisión es capaz de capturar interacciones, ofrece una buena robustez frente al ruido y proporciona una comparación clara frente a modelos de boosting más avanzados.(62)

Modelo XGBoost

XGBoost es un modelo de boosting que mejora las predicciones a partir de errores previos. A diferencia del modelo Random Forest, cada árbol subsiguiente aprende a partir de los árboles anteriores y no tiene asignado el mismo peso (63)

Modelo CatBoost

El modelo CatBoost(64) tiene como característica que utiliza las variables categóricas originales sin tratamiento es decir sin realizar modificación a one hot ni normalizar o estandarizar las variables. También utiliza árboles, pero estos son simétricos (65)

Resultados de los modelos

Se entrena los modelos con validación cruzada y 5 folds para asegurar que se entrena los modelos con todos los datos. No se ajustan hiperparámetros pudiendo los modelos mejorar en cuanto a rendimiento si se hubieran ajustado. Se crea una tabla resumen donde se indica el nombre del modelo, el RMSE medio el RMSE standard de cada fold, el RMSE medio y mediano de precio relativo (valor del error RMSE / precio medio, valor del error RMSE / precio mediano) el tiempo de ejecución en segundos y los parámetros utilizados para cada modelo

	model	rmse_mean	rmse_std	rmse_rel_mean	rmse_rel_median	time_sec	params
0	Ridge	232,189.96	236,866.12	0.520	0.580	0.167	{'alpha': 1.0}
1	RandomForest	88,711.31	4,305.16	0.199	0.222	10.342	{'n_estimators': 500, 'max_depth': None, 'min_samples_leaf': 2, 'random_state': 42, 'n_jobs': -1}
2	XGBoost	84,328.22	3,614.18	0.189	0.211	2.242	{'n_estimators': 500, 'learning_rate': 0.05, 'max_depth': 6, 'subsample': 0.8, 'colsample_bytree': 0.8, 'objective': 'reg:squarederror', 'random_state': 42, 'n_jobs': -1}
3	CatBoost	84,223.25	3,459.15	0.188	0.211	809.979	{'iterations': 1000, 'learning_rate': 0.05, 'depth': 8, 'loss_function': 'RMSE', 'early_stopping_rounds': 30, 'random_seed': 42, 'verbose': 0}

Tabla 19 Resultado modelos de predicción precio vivienda

Comparación de modelos (RMSE relativo sobre precio medio)

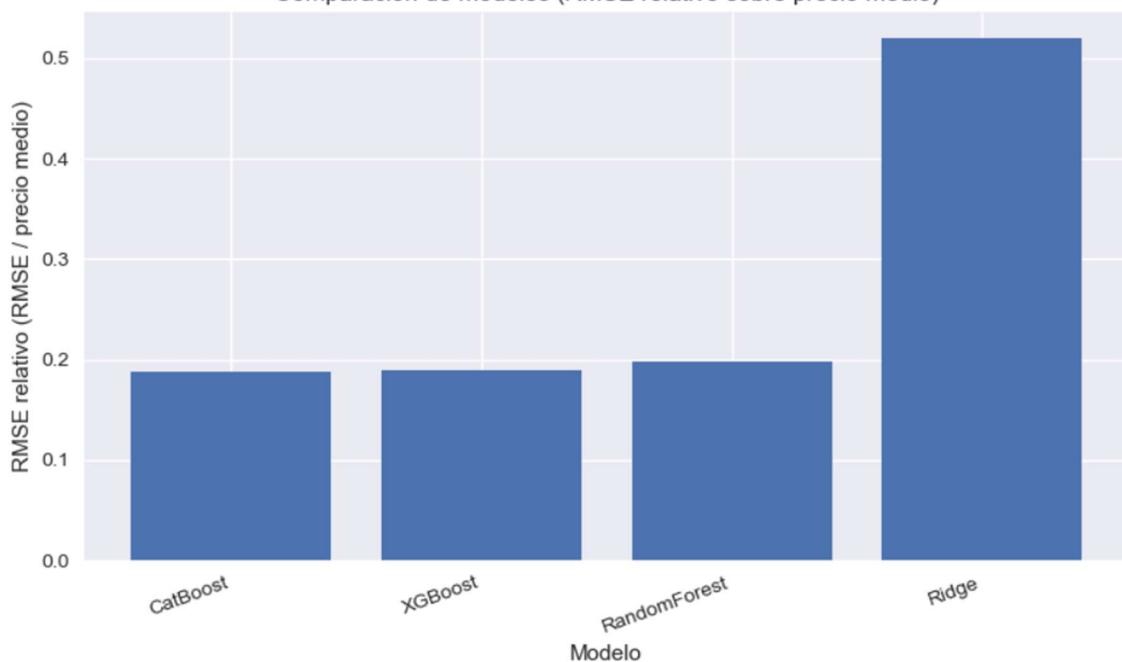


Figura 30 Comparación error relativo sobre precio medio de los modelos

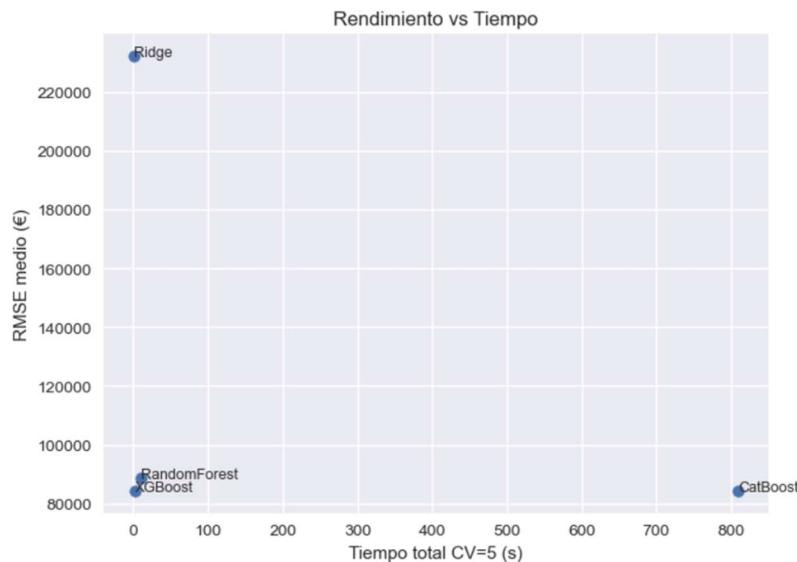


Figura 31 Relación error RMSE con el tiempo de ejecución

En la tabla de resultados, así como en las gráficas se observa el que el modelo Ridge es el que peor desempeño tiene en cuanto a RMSE y el que mejor desempeño tiene en cuanto a tiempo de ejecución. El modelo Random Forest mejora los resultados de Ridge obteniendo una media de RMSE de 88711.31 e incrementando el tiempo de ejecución. El modelo XGBoost mejora mucho a los dos modelos anteriores en rendimiento con una media de RMSE de 84328.22, un error RMSE_STD de 3614.18 y un RMSE relativo medio de 0.189 en 2.224 segundos. CatBoost es el modelo que mejor RMSE medio tiene con 84.223.25 euros y un y un RMSE relativo medio de 0.188 sin embargo es el que más tarda de todos con 809.979 segundos. Como conclusiones se decide escoger los dos modelos con mejor desempeño XGBoost y CatBoost entrenar el modelo sin cross validation y comparar los resultados para decidir qué modelo se comporta mejor con el conjunto de datos.

Decisión de mejor modelo

Para determinar el mejor modelo se calcula el RMSE en diferentes tramos de precio a fin de poder seleccionar un modelo u otro en función del error ya que uno de ellos puede funcionar muy bien para precios bajos y otro para precios altos dadas las características del dataset. Se entrena los dos modelos realizando una separación del 20% de los datos para el test y el resto para entrenamiento. Se observa como el modelo CatBoost obtiene mejores resultados en los tramos bajo, medio-bajo y medio-alto el 75% del conjunto de datos y XGBoost obtiene mejores resultados en el tramo Alto. El modelo CatBoost tiene error porcentual medio MAPE de 14.38%. El tiempo que tarda adicional a XGBoost se debe al

tratamiento de las variables ya que no necesita normalización o estandarización. Se escoge el modelo CatBoost debido a su desempeño total.

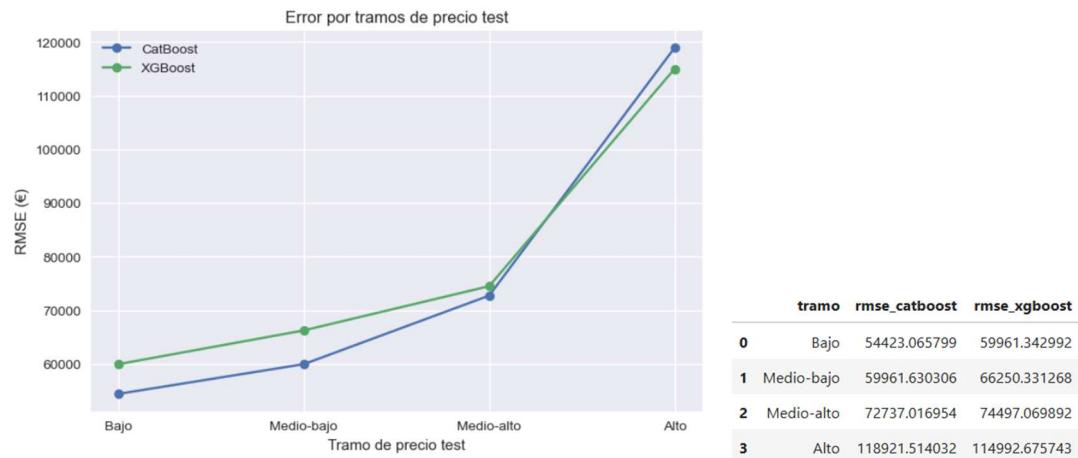


Figura 32 Gráfica error del modelo por tramos de precio

Las variables que más influyen al modelo son el tamaño de la vivienda con un 28.80% seguida de la latitud, el número de baños, el barrio y el distrito.

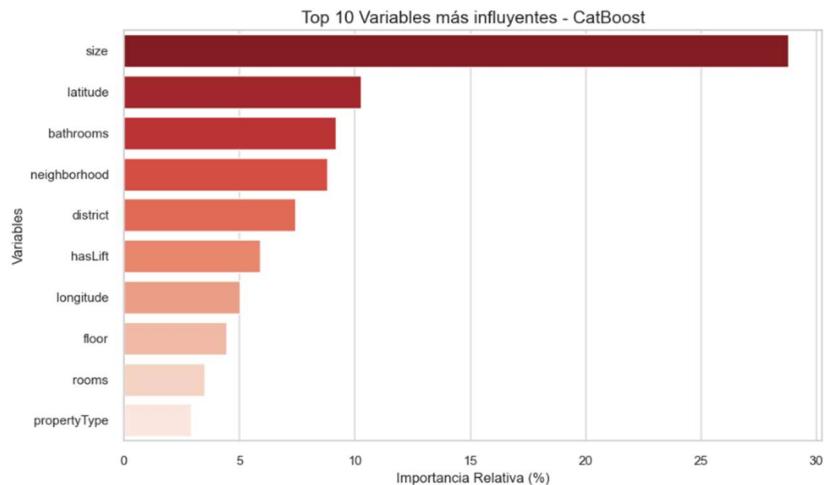


Figura 33 Gráfica de variables más influyentes modelo CatBoost

Resumen de resultados

Se ha conseguido realizar una clasificación de los distritos en base a datos sociodemográficos obteniendo una descripción detallada que permite a personas que no conocen la ciudad poder determinar en qué distrito establecer su residencia. Se han obtenido 3 grupos de distritos: clúster 0 distritos de clase alta, acomodados, clúster 1 distritos conflictivos con presión urbana o turística y clase trabajadora y clúster 2 distritos residenciales y asequibles, baja renta y nivel estudios. Observa la relación entre valoración media y valor real destaca Nou Barris y Sant Andreu con precio bajo y alta valoración.



Figura 34 Valor percibido frente valor precio vivienda en portal idealista etiquetas distrito

Se ha obtenido un modelo capaz de predecir el precio de una vivienda en función de sus características teniendo en cuenta los datos sociodemográficos de los distritos. Se ha creado una visualización en la que el usuario puede seleccionar tipo de vivienda sea alquiler de habitación, alquiler o compra, precio características de vivienda y puede indicar la importancia de la movilidad en servicios públicos, seguridad, limpieza... La aplicación devuelve los distritos más coincidentes, así como un número de viviendas destacadas, puede comparar distritos descubrir las características de cada uno y obtener la predicción de precio de vivienda del modelo creado introduciendo los datos requeridos de la misma.

4. Conclusiones y trabajos futuros

Conclusiones

Con el trabajo final se ha conseguido realizar un análisis del mercado inmobiliario en el área metropolitana de Barcelona, analizando los 10 distritos en cuanto a nivel de satisfacción de los residentes como con datos oficiales de servicios de movilidad y criminalidad.

Conclusiones técnicas: Se han obtenido 5 componentes principales de la valoración sociodemográfica de la población, tras realizar el PCA y Kmeans, se obtienen 3 clúster de distritos: clúster 0, distritos de clase alta, acomodados, clúster 1, distritos conflictivos con presión urbana o turística de clase trabajadora y clúster 2, distritos residenciales y asequibles, baja renta y nivel estudios. Se ha conseguido realizar una predicción de precios de vivienda en base a datos actualizados del portal Idealista obteniendo dos modelos de predicción de precios destacando el modelo CatBoost con RMSE medio 84223.25 euros y

un RMSE relativo medio de 0.188 con un error porcentual medio de 14.38% teniendo mejor desempeño que el modelo XGBoost en un 75 por ciento de los datos analizados por tramos. También se han obtenido las variables más influyentes al modelo en cuanto al precio de la vivienda siendo el tamaño de la vivienda 28.8% la más influyente junto con la latitud 10.26%, número de baños 9.17%, barrio 8.81% y distrito 7.43%.

Conclusiones metodológicas: Al inicio del trimestre se tuvo que activar la tabla de gestión de riesgos debido a un fallo en el disco duro del ordenador. Esto manifestó el error de no haber realizado copia de seguridad recurrente perdiendo todos los datos y debiendo de comenzar de nuevo. En cuando al EDA se ha logrado realizar un buen trabajo, extrayendo y limpiando datos de repositorios de datos abiertos, obteniendo nuevos datos mediante feature engineering. Se han creado nuevos conjuntos de datos (ver Anexo) que han permitido conseguir el análisis del mercado y predicción de precios. Se ha conseguido desarrollar una visualización en Streamlit (enlace en anexo) suponiendo un reto de aprendizaje en cuanto a diseño de usuario UX. Se ha logrado realizar una clasificación de distritos agrupándolos y mostrando tanto las similitudes como las diferencias entre ellos.

Conclusiones de negocio: Si bien el trabajo final se enmarca dentro del ámbito académico y no se plantea seguir la línea de negocio, puede ser un punto de partida para desarrollar la aplicación o para nuevos trabajos que quieran continuar con este proyecto. Se ha creado una visualización la cual permite al usuario descubrir el barrio más afín a sus intereses, así como realizar comparación de distritos, obtener viviendas recomendadas y obtener un precio dadas las características de una vivienda. Se han cruzado datos permitiendo detectar oportunidades de inversión, así como relacionar la calidad de vida con el precio de mercado en las ofertas publicadas en Idealista. Se han obtenido las variables que más influyen en el precio de la vivienda. Sin embargo, el trabajo tiene limitaciones, la primera el tiempo de desarrollo del trabajo ya que se encuentra dentro del ámbito académico y se desarrolla dentro de los créditos de la asignatura. La segunda son los datos, ya que solo se han analizado publicaciones recientes en el portal de Idealista y en un tamaño reducido debido a la limitación mensual de la API. Los objetivos no conseguidos en este trabajo son:

Recomendar vivienda en función de puntos de interés y zona de trabajo, así como el cálculo de rutas y tiempo de desplazamiento con acceso a aplicación tipo Maps el cual no se ha llevado a cabo debido a la carga que ha supuesto el trabajo de ETL como el de EDA. El fine tuning de los modelos para obtener mejores resultados no se ha podido realizar debido al tiempo disponible y a la falta de datos para poder realizar las pruebas necesarias.

Trabajos futuros

Este trabajo puede ser el punto de partida de un análisis mucho más profundo o incluso de una aplicación funcional basada en el análisis del mercado inmobiliario en el Área Metropolitana de Barcelona. Para ello se describen las posibles líneas de trabajo futuras:

Línea de trabajo a corto plazo: Obtener más datos actualizados de la API de Idealista que permitan realizar el fine tuning de los modelos, realizar las búsquedas de recomendación de vivienda en tiempo real, alimentar el modelo con datos históricos, desplegar un modelo para las habitaciones y viviendas en alquiler. Obtener datos a nivel barrio o calle para ajustar más la predicción del modelo y determinar si una finca es buena o mala inversión.

Línea de trabajo a medio plazo: Obtener series temporales de datos de vivienda para poder realizar una predicción de precio a futuro, alimentar el modelo con más datos sociodemográficos como población en activo, paro, inmigración, IPC.

Línea de trabajo a largo plazo: Obtener precios de venta real (Portal Inmobiliario de Notariado), incorporar datos de hipotecas concedidas, amortizaciones y gastos, determinar la rentabilidad en caso de compra de vivienda y obtener un beneficio alquilando una vivienda.

El trabajo final de máster ha supuesto un desafío mucho mayor al realizado en el trabajo final de grado, debido al tiempo disponible para planificar, diseñar y desarrollar. Se ha desarrollado un análisis integral del mercado inmobiliario en el área metropolitana de Barcelona combinando información socioeconómica, datos de movilidad, seguridad y percepción ciudadana a nivel de distrito junto con datos de oferta real en portal inmobiliario Idealista. Mediante técnicas de reducción de dimensionalidad (PCA), algoritmo no supervisado de segmentación (Kmeans) y modelos predictivos lineares, de árboles y boosting (Ridge, Random Forest, XGBoost, CatBoost) se ha conseguido predecir el precio de vivienda aportando información adicional de la percepción ciudadana y datos oficiales así como una relación entre el valor de mercado y el valor percibido por la población que permite detectar distritos que contengan viviendas que sean una oportunidad de mercado con precios inferiores pero también detectar viviendas con precios fuera de mercado. Finalmente, todo el trabajo se ve reflejado en la visualización creada permitiendo realizar recomendaciones al usuario en función de sus necesidades e intereses y mostrando los distritos de manera que se puedan comparar en función de sus características.

Este trabajo muestra como la ciencia de datos permite obtener insights relevantes de datos reales dispersos facilitando el conocimiento y la toma de decisiones basada en datos.

5. Glosario

- Boxplot:** Diagrama de caja que resume la distribución de variable numérica.
- Componente principal:** Resultado de PCA contiene la información de variables originales
- Correlación:** Indica la relación entre dos variables
- Clúster:** Grupo de observaciones similares
- Data Leakage:** Fuga de datos, el modelo se entrena con datos que contienen información de la variable objetivo.
- Dataset:** Conjunto de datos.
- Distrito:** División administrativa que agrupa barrios.
- Early Stopping:** Detiene el entrenamiento del modelo cuando el rendimiento no mejora.
- EDA** Exploratory Data Analysis: Exploración inicial que incluye limpieza de datos
- Fold:** Partición de datos en validación cruzada.
- Kmeans:** Algoritmo de agrupación que agrupa en un número k de clústeres
- Mockup:** Representación funcional no finalizada
- NaN:** Not a Number, representa valor faltante.
- Outliers:** Valores atípicos
- PCA:** Análisis de componentes principales: Técnica de reducción de dimensionalidad transforma variables en componentes principales.
- RMSE:** Error promedio entre valores reales y predichos.
- Validación cruzada - Cross Validation:** Divide el dataset en subconjuntos folds.que son entrenados iterativamente hasta haber entrenado el número de folds seleccionado.

6. Bibliografía

1. Home - United Nations Sustainable Development [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://www.un.org/sustainabledevelopment/>
2. GanttPRO [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://app.ganttpro.com/#/project/1718524373015/gantt>
3. Trema Multimedia, FUOC. La presentación del Trabajo Final de Grado – 1 [Internet]. 2020 [citado 4 de noviembre de 2025]. Disponible en: https://materials.campus.uoc.edu/cdocent/PID_00275762/
4. Trema Multimedia, FUOC. La presentación del Trabajo Final de Grado – 2 [Internet]. 2020 [citado 4 de noviembre de 2025]. Disponible en: https://materials.campus.uoc.edu/cdocent/PID_00275760/
5. Consultas frecuentes - Cliente Bancario, Banco de España [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://clientebanco.bde.es/pcb/es/menu-horizontal/podemosayudarte/consultasreclama/consultasreclama/>
6. BOE-A-2018-16673 Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales. [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://www.boe.es/buscar/act.php?id=BOE-A-2018-16673>
7. Evolución del precio de la vivienda en alquiler en España — idealista [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://www.idealista.com/sala-de-prensa/informes-precio-vivienda/alquiler/>
8. Informes y estudios del mercado inmobiliario - idealista/data [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://www.idealista.com/data/estudios-de-mercado/>
9. Guía para vivir en Barcelona: Precios y zonas - yaencontre [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://www.yaencontre.com/guias/barcelona>
10. Sede Electrónica del Catastro - Inicio [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://www.sedecatastro.gob.es/>
11. Portal de la Dirección General del Catastro: Preguntas frecuentes [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://www.catastro.hacienda.gob.es/esp/faqs.asp>
12. Portal Estadístico del Notariado | Inmobiliario [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://penotariado.com/inmobiliario/>
13. García Benítez EJ, Casar JG. Big data y analítica de datos en el sector de alquiler de viviendas. 2024 [citado 4 de noviembre de 2025]; Disponible en: <https://repositorio.unican.es/xmlui/handle/10902/35648>
14. inmobiliario M, Carlos Casas del rosal J, Casas del rosal david, María Caridad ocerin Julia Núñez Tabales J. Mercado inmobiliario de España: Una herramienta para el análisis de la oferta. Revista de Economía y Finanzas [Internet]. 12 de febrero de 2019 [citado 4 de noviembre de 2025];42(120). Disponible en: <https://reveyf.es/index.php/REyF/article/view/82>

15. Álvarez Martín T. Análisis y predicción del mercado inmobiliario en la Comunidad de Madrid [Internet]. Facultad de Estudios Estadísticos (UCM); 2018 [citado 4 de noviembre de 2025]. Disponible en: <https://hdl.handle.net/20.500.14352/14240>
16. Open Data BCN | Servicio de datos abiertos del Ajuntament de Barcelona [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://opendata-ajuntament.barcelona.cat/es/>
17. Smou. Muévete fàcil, muévete mejor | SMOU [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://www.smou.cat/es>
18. Marmolejo-Duarte C, Espinoza-Zambrano P, Biere-Arenas R. Changes in the effect of energy efficiency, centrality and architectural quality on multi-family values in Barcelona 2020- 2023. Ciudad y Territorio Estudios Territoriales. 18 de diciembre de 2024;56(222):1283-306.
19. ¿Qué es EDA (Exploratory Data Analysis) en Data Science? - NDS [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://nuclio.school/blog/eda-exploratory-data-analysis/>
20. Definición de protección de datos - Diccionario panhispánico del español jurídico - RAE [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://dpej.rae.es/lema/protecci%C3%B3n-de-datos>
21. Protección de Datos conforme al reglamento RGPD - Your Europe [Internet]. [citado 4 de noviembre de 2025]. Disponible en: https://europa.eu/youreurope/business/dealing-with-customers/data-protection/data-protection-gdpr/index_es.htm#inline-nav-3
22. 3 Ways to Embrace Proactive Data Ethics [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://www.gartner.com/smarterwithgartner/3-ways-to-embrace-proactive-data-ethics>
23. La ética en la gestión de los datos | datos.gob.es [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://datos.gob.es/es/noticia/la-etica-en-la-gestion-de-los-datos>
24. The Data Ethics Canvas | The ODI [Internet]. citado 4 de noviembre de 2025]. Disponible en: <https://theodi.org/insights/tools/the-data-ethics-canvas-2021/>
25. Reglamento - 2016/679 - EN - GDPR - EUR-Lex [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=celex%3A32016R0679>
26. ¿Qué es el análisis de clientes? | IBM [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://www.ibm.com/es-es/think/topics/customer-analytics>
27. Customer Analytics | Deloitte España [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://www.deloitte.com/es/es/services/consulting/services/customer-analytics.html>
28. Los juegos que ayudaron a la IA a evolucionar | IBM [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://www.ibm.com/history/early-games>
29. ¿Qué es el machine learning y que usos tiene? | Repsol [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://www.repsol.com/es/energia-futuro/tecnologia-innovacion/machine->

- learning/index.cshtml?gad_source=1&gclid=Cj0KCQiA4L67BhDUARIsADWrl7G3csgS4vDcjGshddS7EPAJLBQOLTSj50GYSJoArA8e9EklhhQ_Q6QaAuzxEALw_wcB
30. ¿Qué es el Aprendizaje mediante refuerzo? - Explicación del Aprendizaje mediante refuerzo - AWS [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://aws.amazon.com/es/what-is/reinforcement-learning/>
31. 2.3. Clustering — scikit-learn 1.6.0 documentation [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://scikit-learn.org/stable/modules/clustering.html#k-means>
32. Métricas en regresión. La regresión es importante y conocer... | by Nicolás Arrioja Landa Cosio | Medium [Internet]. [citado 5 de noviembre de 2025]. Disponible en: <https://medium.com/@nicolasarioja/m%C3%A9tricas-en-regresi%C3%B3n-%C3%B3n-5e5d4259430b>
33. Análisis multivariante. Regresión lineal múltiple - Evidencias en pediatría [Internet]. [citado 5 de noviembre de 2025]. Disponible en: <https://evidenciasenpediatria.es/articulo/8192/analisis-multivariante-regresion-lineal-multiple>
34. RandomForestClassifier — scikit-learn 1.7.2 documentation [Internet]. [citado 5 de noviembre de 2025]. Disponible en: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
35. XGBoost Documentation — xgboost 3.1.1 documentation [Internet]. [citado 5 de noviembre de 2025]. Disponible en: <https://xgboost.readthedocs.io/en/stable/>
36. ¿Qué es una RNN?: Explicación sobre redes neuronales recurrentes: AWS [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://aws.amazon.com/es/what-is/recurrent-neural-network/>
37. Our Documentation | Python.org [Internet]. [citado 5 de noviembre de 2025]. Disponible en: <https://www.python.org/doc/>
38. Jupyter Notebook Guide | Databricks [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://www.databricks.com/glossary/jupyter-notebook>
39. Matplotlib documentation — Matplotlib 3.10.7 documentation [Internet]. [citado 5 de noviembre de 2025]. Disponible en: <https://matplotlib.org/stable/>
40. Waskom M. seaborn: statistical data visualization. J Open Source Softw. 6 de abril de 2021;6(60):3021.
41. Flourish [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://flourish.studio/>
42. Streamlit documentation [Internet]. [citado 5 de noviembre de 2025]. Disponible en: <https://docs.streamlit.io/>
43. Gradio Documentation [Internet]. [citado 5 de noviembre de 2025]. Disponible en: <https://www.gradio.app/docs>
44. Software de análisis e inteligencia de negocios | Tableau [Internet]. [citado 4 de noviembre de 2025]. Disponible en: <https://www.tableau.com/es-es>
45. Población | Barcelona Datos | Ajuntament de Barcelona [Internet]. [citado 14 de diciembre de 2025]. Disponible en: <https://portaldades.ajuntament.barcelona.cat/es/estad%C3%ADsticas/yzlntdm2fs>

46. Superficie de los barrios de la ciudad de Barcelona | Barcelona Datos | Ajuntament de Barcelona [Internet]. [citado 14 de diciembre de 2025]. Disponible en: <https://portaldades.ajuntament.barcelona.cat/es/microdatos/8f144d2c-1185-4e5c-9b97-ac930eeffca8>
47. Número de hechos delictivos conocidos por la Policía de la Generalitat de Catalunya-Mossos d'Esquadra por categoría y tipología | Barcelona Datos | Ajuntament de Barcelona [Internet]. [citado 12 de diciembre de 2025]. Disponible en: <https://portaldades.ajuntament.barcelona.cat/es/estad%C3%ADsticas/ocbk1bftni?view=table>
48. Listado de equipamientos de transportes y servicios relacionados de la ciudad de Barcelona | Barcelona Datos | Ajuntament de Barcelona [Internet]. [citado 14 de diciembre de 2025]. Disponible en: <https://portaldades.ajuntament.barcelona.cat/es/microdatos/d55f001a-d105-4095-84db-fa69514b84ba>
49. Estaciones de autobus de la ciudad de Barcelona | Barcelona Datos | Ajuntament de Barcelona [Internet]. [citado 14 de diciembre de 2025]. Disponible en: <https://portaldades.ajuntament.barcelona.cat/es/microdatos/d395e808-697d-4722-8eb9-b672a8ba0916>
50. Equipamientos culturales de la ciudad de Barcelona - Conjuntos de datos - Open Data Barcelona [Internet]. [citado 11 de diciembre de 2025]. Disponible en: <https://opendata-ajuntament.barcelona.cat/data/es/dataset/equipaments-culturals-icub>
51. Encuesta de Servicios Municipales de la ciudad de Barcelona - Evolución - Conjuntos de datos - Open Data Barcelona [Internet]. [citado 11 de diciembre de 2025]. Disponible en: <https://opendata-ajuntament.barcelona.cat/data/es/dataset/esm-bcn-evo>
52. Evolución del precio de la vivienda en venta en Barcelona — idealista [Internet]. [citado 12 de diciembre de 2025]. Disponible en: <https://www.idealista.com/sala-de-prensa/informes-precio-vivienda/venta/cataluna/barcelona-provincia/barcelona/>
53. Vulnerabilidad social [Internet]. [citado 12 de diciembre de 2025]. Disponible en: <https://www.miteco.gob.es/es/cambio-climatico/temas/impactos-vulnerabilidad-y-adaptacion/plan-nacional-adaptacion-cambio-climatico/vuln-social.html>
54. ¿Qué es el análisis de componentes principales (PCA)? | IBM [Internet]. [citado 14 de diciembre de 2025]. Disponible en: <https://www.ibm.com/es-es/think/topics/principal-component-analysis>
55. Scree Plot. Principal Component Analysis (PCA) is a... | by SANCHITA MANGALE | Medium [Internet]. [citado 14 de diciembre de 2025]. Disponible en: <https://sanchitamangale12.medium.com/scree-plot-733ed72c8608>
56. Request API access [Internet]. [citado 21 de diciembre de 2025]. Disponible en: <https://developers.idealista.com/access-request>
57. 406 Not Acceptable - HTTP | MDN [Internet]. [citado 21 de diciembre de 2025]. Disponible en: <https://developer.mozilla.org/en-US/docs/Web/HTTP/Reference>Status/406>

58. ¿Qué es un comando Curl y cómo usarlo? [Internet]. [citado 21 de diciembre de 2025]. Disponible en: <https://www.hostinger.com/es/tutoriales/comando-curl>
59. ¿Qué es la fuga de datos en el machine learning? | IBM [Internet]. [citado 24 de diciembre de 2025]. Disponible en: <https://www.ibm.com/es-es/think/topics/data-leakage-machine-learning>
60. Qué es la regresión Lasso | IBM [Internet]. [citado 25 de diciembre de 2025]. Disponible en: <https://www.ibm.com/es-es/think/topics/lasso-regression>
61. ¿Qué es la regresión Ridge? | IBM [Internet]. [citado 25 de diciembre de 2025]. Disponible en: <https://www.ibm.com/es-es/think/topics/ridge-regression>
62. ¿Qué es un bosque aleatorio? | IBM [Internet]. [citado 25 de diciembre de 2025]. Disponible en: <https://www.ibm.com/es-es/think/topics/random-forest>
63. Cómo funciona el algoritmo XGBoost—ArcGIS Pro | Documentación [Internet]. [citado 25 de diciembre de 2025]. Disponible en: <https://pro.arcgis.com/es/pro-app/3.3/tool-reference/geoai/how-xgboost-works.htm>
64. CatBoost - open-source gradient boosting library [Internet]. [citado 25 de diciembre de 2025]. Disponible en: <https://catboost.ai/>
65. Cómo funciona el algoritmo CatBoost—ArcGIS Pro | Documentación [Internet]. [citado 25 de diciembre de 2025]. Disponible en: <https://pro.arcgis.com/es/pro-app/3.3/tool-reference/geoai/how-catboost-works.htm>

7. Anexos

Arquitectura de datos y modelado analítico del mercado inmobiliario de Barcelona

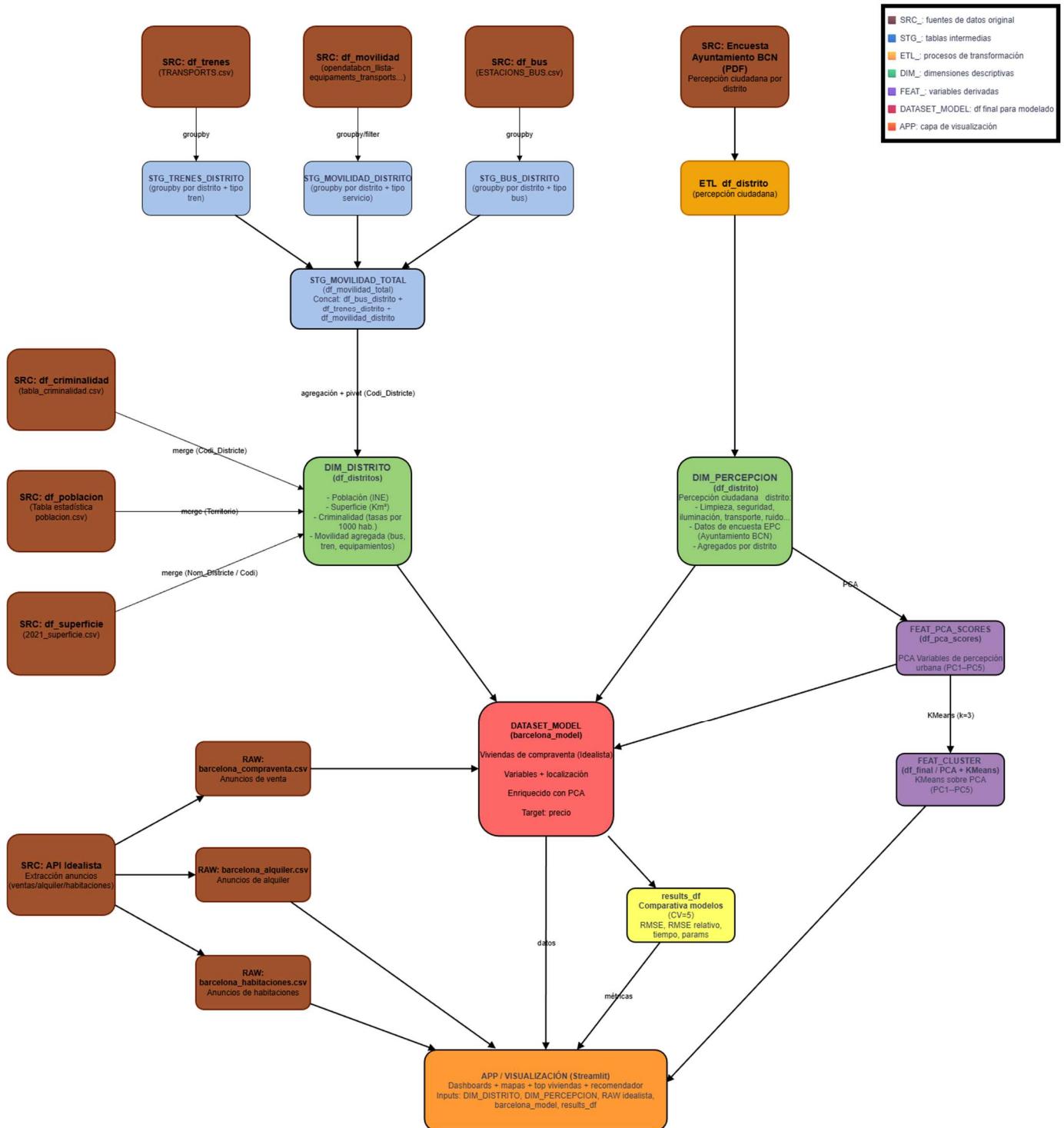


Figura 35 Arquitectura de datos y modelado

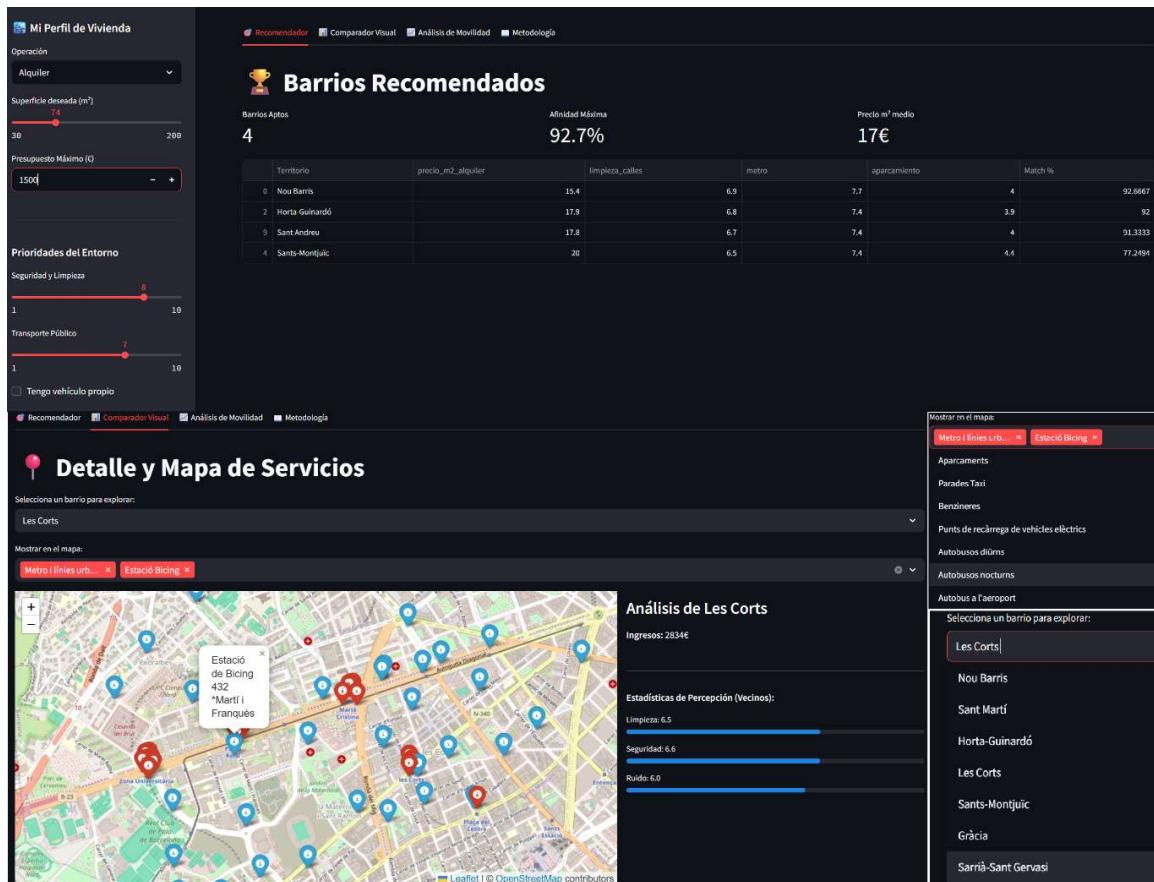


Figura 36 Mockup de visualización (boceto)

Aplicación creada en Streamlit

El proyecto ha sido desplegado como una aplicación web mediante Streamlit Community Cloud, a partir de un repositorio público en GitHub.

El repositorio incluye el código fuente completo, el modelo predictivo entrenado (CatBoost) y los conjuntos de datos utilizados por la aplicación. La aplicación permite la interacción con los datos y la ejecución del modelo de predicción en tiempo real. Debido a las limitaciones de Streamlit se recomienda el uso en ordenador.

Repositorio GitHub: <https://github.com/Abelibz/tfm-streamlit-app>

Aplicación web: <https://tfm-abel-mora-vazquez.streamlit.app/>